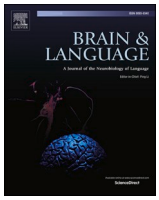
Brain & Language 277 (2026) 105752

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l



Advancing ERP-based tools for language monitoring in chronic aphasia: psychometric and clinical utility of the word-level N400

Sarah Grace Dalton [a,*], Mark Lavelle [b], James F. Cavanagh [b], Jessica D. Richardson [c,d,e]

[a] Department of Communication Sciences and Special Education, University of Georgia, 110 Carlton St, Athens, GA 30602 USA
[b] Department of Psychology, University of New Mexico, MSC02, Albuquerque, NM 87131, USA
[c] Department of Speech and Hearing Sciences, University of New Mexico, MSC01, Albuquerque, NM 87131, USA
[d] Center for Brain Recovery and Repair, University of New Mexico School of Medicine, MSC08 4720, Albuquerque, NM 87131, USA
[e] New Mexico Alzheimer's Disease Research Center, 1101 Yale NE, Albuquerque, NM 87106, USA

ARTICLE INFO

Keywords:
Aphasia
ERP
N400
Psychometrics

ABSTRACT

The N400 is a potential biomarker for cognitive-linguistic recovery in people with aphasia (PWA), as several studies have demonstrated treatment-induced changes. However, the validity, reliability, and functional significance of the N400 as a repeated measure in this population remains unverified, whether elicited through sentence- or word-level paradigms. This study addresses these key questions required for the N400 to be utilized as a biomarker of recovery in PWA. First, we developed and validated word-level semantic and phonological tasks to elicit the N400, compared group responses between healthy controls and PWA, examined the reliability and stability of N400 across groups, and explored relationships between the N400 and behavioral language performance. Our findings demonstrate successful validation of semantic and phonological tasks that elicit the N400, and two control tasks that do not. These tasks elicited N400 responses in PWA and controls, indicating similar underlying processing, albeit with greater variability in PWA. Reliability and stability of the N400 were mixed, indicating that caution is warranted when using N400 as a treatment response measure. Critically, condition-specific N400 amplitudes were significantly associated with functional language measures, highlighting the behavioral relevance of this neurophysiological signal. Together, these results provide validated elicitation tools, clarify limits on N400 reliability, and identify meaningful brain-behavior relationships, advancing the N400 as a more precise and interpretable biomarker for tracking recovery and treatment outcomes in aphasia.

1. Introduction

1.1. The history of N400 research in aphasia

In this study, we investigate the validity and reliability of using the N400 as a measure of language processing in people with aphasia (PWA). The N400 is a cognitive-linguistic ERP component frequently used to study language processing in PWA following a stroke or other brain injury. PWA exhibit distinct differences in their N400 responses, characterized by reduced amplitude, delayed onset, and more varied distribution patterns across electrode sites (Arheix-Parras et al., 2023, Meechan et al., 2021). These findings suggest that, as a group, PWA have compromised yet intact access to lexical, semantic, and/or phonological information. For example, individuals with more severe auditory comprehension impairments show significant differences in N400 responses compared to controls, while those with higher comprehension abilities do not display the same differences (Arheix-Parras et al., 2023, Meechan et al., 2021).

EEG has also been increasingly used to track neural responses to various aphasia treatments, including behavioral therapy, noninvasive brain stimulation, and pharmacological interventions (Arheix-Parras et al., 2023, Meechan et al., 2021; Silkes & Anjum, 2021). While promising, definitive conclusions about treatment responses as they relate to EEG metrics, including the N400, remain elusive. As revealed in systematic reviews (Arheix-Parras et al., 2023, Meechan et al., 2021; Silkes & Anjum, 2021), in many cases, increased N400 amplitude is

* Corresponding author.
E-mail addresses: sgdalton@uga.edu (S.G. Dalton), marklavelle13@unm.edu (M. Lavelle), jcavanagh@unm.edu (J.F. Cavanagh), jdrichardson@unm.edu (J.D. Richardson).

https://doi.org/10.1016/j.bandl.2026.105752
Received 18 July 2025; Received in revised form 11 March 2026; Accepted 1 April 2026
Available online 9 April 2026

associated with recovery, treatment improvements, and/or the intensity of practice or treatment schedules (e.g., Aerts et al., 2015; Barbancho et al., 2015; Barwood et al., 2011; Barwood et al., 2012). Conversely, some studies (though fewer in number) report that decreased amplitude is linked to improvement (e.g., Aerts et al., 2015; Barbancho et al., 2015). Findings regarding latency and topographic distribution are sparse, with only one study specifically addressing distribution patterns. Only three studies (from one research group) have explored the relationship between lesion location and the N400 (Meechan et al., 2021).

1.2. Gaps and needs in N400 research in aphasia

The N400 in PWA has unique potential as a sensitive biomarker for both the presence and severity of aphasia, as well as recovery and treatment responsivity. It is a promising tool for tracking neural activation changes that could lead to behavioral improvements, supporting the case for continued therapy in PWA "on the cusp" of responding to treatment. However, the clinical utility of the N400 is still limited. For example, some PWA remain unable to participate in tasks that reliably elicit the N400, as EEG analysis requires participants to be attentive, engaged, and mostly *accurate* in their responses. This means that we currently know very little about N400 responses in people with more severe aphasia. On the other hand, its potential also remains underdeveloped as a biomarker for people with very mild presentations of aphasia, who are often more difficult to distinguish from controls. Enhancing the N400 assessment and/or analyses to accommodate a broader range of aphasia profiles could significantly improve its utility in tracking recovery and treatment responses. This is especially important since even very mild aphasia can limit the ability to return to work and maintain relationships, severely affecting quality of life.

Further development of the N400 as a clinically useful biomarker is hindered by the substantial methodological heterogeneity across the aphasia literature. Methodological details are reviewed and discussed in depth in *Supplementary Section "Methodological Review"* and *Supplementary Table 1*; we summarize the review findings here. Studies differ widely in the stimuli used to elicit the N400, ranging from sentence- and single-word-level paradigms to picture-based approaches, as well as in presentation modality, with auditory, visual, and multimodal formats represented. Task demands vary considerably, some studies use passive reading or listening, whereas others rely on active judgments (e.g., lexical decision, semantic relatedness, etc.). Substantial variability also exists in participant sampling: studies often include small, heterogenous samples (and frequently draw from overlapping cohorts) or restrict inclusion to particular aphasia types or lesion profiles, limiting generalizability. Additionally, N400 identification and analysis procedures vary widely in the time windows used to define the component, with limited justification for parameter choices. While this methodological diversity demonstrates that PWA of many different subtypes can show N400 effects under many different conditions, it means that there are few studies that align across major methodological dimensions, complicating interpretation and comparison across studies. Collectively, the variability in paradigms, participant sampling, and analytic procedures has been cited repeatedly as an obstacle to clinical translation (Arheix-Parras et al., 2023; Meechan et al., 2021; Silkes & Anjum, 2021), underscoring the need for studies using transparent, standardized methods across diverse PWA.

1.2.1. A need to focus on the reliability, stability, and validity of the N400 in healthy and clinical populations

A critical aspect of any clinical application is the reliability and validity of outcome measurement. To date, these variables have received relatively little attention, despite multiple studies using the N400 response as an outcome measure for treatment-induced change in PWA (e.g., Aerts et al., 2015; Barbancho et al., 2015; Barwood et al., 2011; Barwood et al., 2012; Wilson et al., 2012). Barbancho and colleagues did examine N400 responses at two times in a group of healthy controls and though they did not report any statistically significant differences, they also did not report a measure of equivalence or test–retest reliability. Meechan et al. (2021) reported that 29/30 articles reviewed demonstrated good N400 reliability. However, they defined reliability as using a paradigm, stimuli, and time window for measurement that would reliably elicit an N400 response, which is more closely related to validity than reliability.

To determine the appropriateness of the N400 as a biomarker of treatment response and/or recovery over time in PWA, it is crucial to first characterize its test–retest reliability and stability. It has been well described that individuals with neurological damage exhibit greater variability in both behavioral testing and neuroimaging than healthy controls (MacDonald et al., 2006). Given the disagreements in the literature regarding test–retest reliability and stability of N400 in healthy controls (see Cocquyt et al., 2023), it is essential to directly examine these factors in patient populations. Without these critical steps, any conclusions drawn about the use of the N400 as a biomarker remain speculative at best. This position is supported by our previous investigation of test–retest reliability of spectral analysis of resting state EEG (Dalton et al., 2021). Reliability in healthy controls and PWA varied depending on the electrodes and spectral band selected, highlighting the need to directly investigate test–retest reliability of the N400 in clinical populations rather than rely upon reports from healthy controls. Even so, very little work has been conducted in this realm. In a small ($N = 16$) group of individuals with schizophrenia, who also completed a word-pair semantic judgment task to elicit an N400 response, test–retest reliability of the N400 effect over one week was poor, as was stability (Boyd et al., 2014). Similarly, Lew et al. (2007) reported poor reliability and stability of N400 variables in a small sample of individuals with a history of traumatic brain injury ($N = 7$) compared to controls ($N = 21$).

In addition to establishing reliability and stability, it is important to validate N400 tasks by showing that PWA exhibit similar patterns of N400 responses as healthy controls. When controls show an N400 effect for a given linguistic manipulation, and PWA show a comparable effect, it indicates that the task engages the intended cognitive process in the clinical population. This type of convergent validation demonstrates that, despite brain injury, the neural signature of language processing remains measurable. This aspect is often overlooked, as clinical studies that include controls typically emphasize only group differences rather than the expected similarities grounded in theoretical constructs.

Validation should also include demonstrating that N400 responses are meaningfully related to clinically relevant behavioral measures, such as those that predict real-world language functioning. For the N400 to serve as a useful biomarker in PWA, it is not sufficient for the component to be reliably elicited; its amplitude or effect size should reflect individual differences in functional language abilities. Very few previous studies have explored relationships with discrete language skills or overall severity (see brief overview above, 1.1.), but the sparse evidence is inconsistent. Linking N400 responses to measures of functional communication would provide critical evidence of criterion ecological validity, further supporting their potential as predictive biomarkers of recovery or treatment response.

1.3. Purpose

To accommodate the wide range of language abilities in PWA, we focused on time-efficient, single-word visual paradigms. These tasks reduce participant burden while providing reliable indices of language processing (Kutas & Federmeier, 2011). We included a semantic task because single words can reliably elicit N400 responses in PWA (e.g., see Supplementary Table 2), and a phonological task, as manipulations such as rhyme judgment evoke N400 in auditory and visual paradigms (e.g., Khateb et al., 2010; Kutas & Federmeier, 2011; Perrin & Garcia-Larrea, 2003; Praamstra & Stegeman, 1993), though this has not yet been tested in PWA. Lexical decision and orthographic tasks served as control tasks, matched for attention and motor demands but not expected to elicit

N400 (e.g., Pulvermüller et al., 2004), allowing us to better isolate semantic and phonological processing. This approach balances efficiency and rigor while maximizing interpretability of N400 responses in a diverse clinical population.

Building on this task framework and given the growing use of EEG to study language changes and treatment response in PWA, we designed the current study to address key gaps in understanding the N400 response in this population. Specifically, we sought to answer three research questions (RQs). **RQ 1**: Do word-level semantic and rhyme judgement tasks elicit expected ERP responses in healthy controls and PWA? What differences/similarities are there between groups? **RQ 2**: How reliable and stable are N400 variables for semantic and phonological tasks in controls and PWA? **RQ 3**: How are N400 variables related to cognitive-linguistic abilities in PWA? To address these questions, we aimed to do the following:

1: Validate four different single-word linguistic tasks as control and N400 elicitation paradigms in healthy controls and PWA. If validated, compare N400 amplitude and effect for between-group differences.
2: Describe the reliability and stability of N400 amplitude and N400 effect for word-level semantic and phonological tasks for healthy controls and PWA.
3: Determine the relationship between N400 amplitude and N400 effect with cognitive-linguistic abilities in PWA.

2. Methods

2.1. Participants

The protocol for this study was approved by the IRB at the University of New Mexico. Participants were recruited from Albuquerque, New Mexico and surrounding regions through advertisements, posted flyers, word of mouth, and a research participant database. Potential participants with a diagnosis of significant psychiatric mood disorders were excluded (including major depressive disorder or generalized anxiety disorder). Persons with self-reported symptoms of mild depression and anxiety but no clinical diagnosis were included (e.g., Finnigan, Wong, & Read, 2016; Gorišek et al., 2016; Song et al., 2015). All participants were right-handed according to the Edinburgh Handedness Inventory (Oldfield, 1971). English was the primary language used by all participants at the time of testing, and had been for many years, assessed via the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian, Blumenfeld, & Kaushanskaya, 2007). See supplementary Fig. 1 for CONSORT diagrams for each participant group.

2.1.1. Healthy controls

Twenty-seven healthy controls consented to participate in the study (Table 1). The mean age for healthy controls was 60.9 years (SD = 15.5) and 19 were female. Mean years of education was 16.8 (SD = 2.5). Participants were screened to ensure no history of neurological disease or injury that might affect brain function.

2.1.2. Participants with aphasia

Twenty-six PWA consented to participate (Table 1). The mean age for PWA was 60.2 years (SD = 16.1) and 7 were female. Mean years of education was 15 (SD = 3.2). Most PWA had experienced a single stroke, however, the number of strokes ranged from 1 – 4. Mean time post-stroke was 48.5 months (SD = 39.9). All PWA were greater than 12 months post-stroke to ensure that spontaneous recovery was not a factor in the reliability analysis. All PWA completed the Western Aphasia Battery – Revised (WAB-R; Kertesz, 2006) to quantify type and severity of aphasia. WAB-R Aphasia Quotient (AQ) ranged from 36.8 to 98.8 (M = 81.2, SD = 18.2). Nine PWA were diagnosed with anomic, five with conduction, one with Wernicke's, two with Broca's, and one with transcortical motor aphasia. In addition, eight participants were diagnosed with latent aphasia: although they scored above the WAB-R cutoff of 93.8 (indicating no current clinical aphasia), they had a prior aphasia diagnosis, ongoing language complaints, and below-average performance on naming and/or discourse assessment (described below).

Table 1
Demographic and neuropsychological test data for healthy controls and persons with aphasia.

	Controls (N = 27) Mean *(SD)* Range	PWA (N = 26) Mean *(SD)* Range
Age	60.9 *(15.5)* 22–88	60.2 *(16.1)* 25–87
Education	16.8 *(2.5)* 12–22	15 *(3.2)* 7–20
Sex	19 Female 8 Male	7 Female 19 Male
Number of Strokes	–	1.4 *(0.85)* 1–4
Time Post Stroke	–	48.5 *(39.9)* 5–183
RBANS	97.2 *(13.1)* 64–120	63.2 *(14.5)* 43–87
WAIS-PC	12.8 *(3.2)* 6–18	9.6 *(3.8)* 2–16
WAB-R AQ	–	81.2 *(18.2)* 36.8–98.8
BNT	–	40.3 *(16.9)* 0–58
MC Composite	–	85.8 *(43.6)* 4–159

RBANS: Repeatable Battery for the Assessment of Neuropsychological Status; WAIS – PC: Weschler Adult Intelligence Scales – Picture Completion; WAB-R AQ: Western Aphasia Battery – Revised Aphasia Quotient; BNT: Boston Naming Test; MC Composite: Main concept composite.

2.2. Behavioral assessment

All participants (PWA and healthy controls) completed a cognitive and linguistic assessment (Table 1). Cognitive function was assessed with the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS; Randolph, 1998) and the Wechsler Adult Intelligence Scales – Picture Completion subtest (WAIS-PC; Wechsler, Coalson, & Raiford, 2008). Additional language measures were administered to PWA only and included the WAB-R (described above), Boston Naming Test (BNT; Kaplan, Goodglass & Weintraub, 2001), and a shortened version of the Discourse Comprehension Test (DCT; Brookshire & Nicholas, 1997). Discourse production was assessed in PWA with the Discourse Production Test (DPT; Fromm, Forbes, Holland, & MacWhinney, 2020), which required participants to tell stories depicted in pictures/picture sequences, retell the story of Cinderella, and describe how to make a peanut butter and jelly sandwich.

Discourse was analyzed using main concept analysis (MCA; Dalton & Richardson, 2019; Nicholas & Brookshire, 1995; Richardson & Dalton, 2016, 2020); which evaluates how well an individual communicates the gist, or essential elements, of a story. MCA was scored using standardized checklists with accompanying normative data developed by our team (Richardson & Dalton, 2016, 2020; Dalton & Richardson, 2019). Briefly, MCA is scored by comparing participant-produced utterances to a standardized checklist of main concepts (MCs) for each task to evaluate the accuracy and completeness of MC production. For example, during a retelling of the Cinderella story, one of the MCs is "Cinderella danced with the prince" (with three essential elements: 'Cinderella', 'danced', and 'with the prince'). If an individual attempted to produce this MC and said, "Cinderella danced", it would be considered *Accurate but Incomplete* (AI), because it is missing the essential element 'with the prince'. If an individual said "Cinderella walked with the prince" it would be considered *Inaccurate* (because they danced, not walked) *but Complete* (IC), because all three essential elements are represented but

with an inaccuracy. A score is assigned for each MC ranging from 0 for *Absent* (AB) to 3 for *Accurate and Complete* (AC). Scores are then summed to yield the MC composite score. Using this system, we can sensitively assess the quality of discourse production in individuals across the impairment spectrum, from healthy controls to profound aphasia (Dalton & Richardson, 2019; Dalton, Hubbard, & Richardson, 2020). For this study we analyzed the MC composite score, number of AC MCs, the number of errored MCs (inaccurate and/or incomplete), and the number of MCs attempted (AC MCs plus errored MCs).

2.3. Behavioral tasks during EEG

During completion of behavioral tasks, participants were seated ~60 cm (~2 feet) from a desktop computer in a darkened room. Seating varied as some individuals used personal wheelchairs. However, all participants indicated they could see and read text on the monitor clearly before beginning. Stimuli were presented one at a time in white text on a black background to reduce visual fatigue and confounding eye movements. Stimuli were presented on a 57 cm [diagonal] Dell LCD monitor. Stimulus words contained an average of 5.2 letters, subtending 15.8 degrees of visual angle horizontally. The average lowercase letter subtended 2.8° vertically, whereas uppercase letters subtended 3.2° vertically.

Participants completed four language tasks: lexical decision, orthographic, semantic, and rhyme (i.e., phonological). Orthographic, semantic, and rhyme tasks were modeled after Spironelli and Angrilli (2006). For these, participants viewed word pairs and judged whether pairs matched for text case (orthographic), meaning (semantic), or rhyme (rhyme). The first word (Word 1) in each word pair was displayed for 3 s, followed by a 1-second fixation screen showing a green "+" to visually cue participants with aphasia to compare the two words. Next, the second word (Word 2) was displayed until a behavioral response was recorded (3 s maximum). Inter-trial fixation was two seconds, and a red "x" was displayed as a reminder that participants should not compare the two words it separated. Participants were asked to press a green key for matches and a red key for non-matches.

The semantic and rhyme tasks were used to investigate the N400 component while the lexical decision and orthographic tasks served as control conditions. Forty matching pairs and 40 non-matching pairs were presented in a randomized order across two ~ 5-minute blocks for each condition. Tasks were completed in fixed order: lexical decision, semantic, orthographic, and rhyme. Throughout the remainder of the article, for the semantic task, we will refer to the *congruent* condition for matching pairs and the *incongruent* condition for non-matching pairs. For the rhyme task, we will refer to the *rhyme* condition for matching pairs and the *nonrhyme* condition for non-matching pairs.

2.3.1. Lexical decision task

A string of letters composing either a single word or nonword was presented on the screen for a maximum of 3 s (or until a participant responded). Stimuli were separated by a 1 s inter-trial fixation, which displayed a red "x". This task was used as a control since it required activation of word meanings, but no priming or comparisons across stimulus items. A list of 40 real words of 1–2 syllables with 3–6 phonemes was created. Frequency per million (Davies, 2008) was used to ensure a relatively wide range of frequencies ($M = 43.15$, $SD = 49.38$). Conversely, words were selected for relatively high concreteness and imageability (concreteness $M = 598.87$, $SD = 22.59$; imageability $M = 592.58$, $SD = 24.57$; Wilson, 1988). Forty nonwords were used, 20 with allowed English orthography (1–2 syllables, 4–6 phonemes), and 20 with disallowed English orthography (ranging from 3 to 6 graphemes).

2.3.2. Orthographic task

Frequency per million of selected words was higher on average for this task ($M = 143.9$, $SD = 271.6$) compared to the other similar tasks, but varied widely. Concreteness ranged from 290 to 645 ($M = 558.2$, $SD = 72.4$) and imageability from 340 to 639 ($M = 567.6$, $SD = 65$). Word length ranged from 1 to 9 phonemes ($M = 4.2$; $SD = 1.4$) with 1–4 syllables ($M = 1.5$; $SD = 0.7$). The orthographic task served as a control because participants simply judged the agreement of case (upper- versus lower-case) between words; deeper processing of phonology or semantics was not required.

2.3.3. Semantic task

To ensure the semantic relatedness of word pairs, the mutual information (MI; Davies, 2008) score ($M = 4.6$; $SD = 1.9$). The lowest MI score was for the word pair "tree – paper", and the highest MI score was for the word pair "thunder – lightning". Frequency per million varied widely ($M = 41.64$, $SD = 49.35$) while concreteness and imageability were high (concreteness $M = 553.07$, $SD = 89.04$; imageability $M = 566.43$, $SD = 63.78$). Word length ranged from 2 to 14 phonemes ($M = 4.5$; $SD = 1.9$) and 1–6 syllables ($M = 1.7$; $SD = 0.9$). All but three words contained fewer than 10 phonemes and fewer than 5 syllables.

2.3.4. Rhyme task

Frequency per million once again encompassed a wide range ($M = 95.1$, $SD = 202.4$). Concreteness ranged from 254 to 670 ($M = 529.4$, $SD = 98.5$) and imageability from 306 to 639 ($M = 543.6$, $SD = 77.9$). Word length ranged from 2 to 8 phonemes ($M = 3.9$; $SD = 1$) and 1–2 syllables ($M = 1.4$; $SD = 0.5$). Words in this task were shorter, as the number of rhyme pairs typically decreases as length increases. When identifying rhyme pairs, care was taken to include words with and without overlapping orthography (e.g., cat/rat versus tail/scale). This step was taken to reduce judgments based solely on shared orthography rather than the underlying phonology. Based on the results, it is unlikely that orthographic similarity alone accounted for responses, given the differential performance across the orthographic and rhyme tasks for both groups.

2.4. EEG acquisition

EEG data were recorded from 64 electrodes placed according to the international 10–10 system (Chatrian et al., 1985) using a stretch-lycra cap. Electrode FPz (located in the middle of the forehead) served as the ground and CPz (located along the midline of the head from left to right, just behind the center of the head from front to back) as the online reference. Two additional channels recorded heart rate and eye movements of the left eye from pairs of bipolar electrodes. (Only one eye was recorded as eye movements generally occur in tandem.) EEG was sampled at 500 Hz on a BrainVision actiCHamp system using Pycorder. Online bandpass filtering from 0.01 Hz to 100 Hz was applied. Data from FT9, FT10, TP9, and TP10 were discarded prior to analyses due to low signal to noise ratio.

2.5. EEG processing

Offline analyses were conducted with custom scripts in MATLAB using the EEGLAB toolbox (v. 2020_0, Delorme & Makeig, 2004). The average reference was calculated and CPz was recovered (pop_reref). A zero-phase, non-causal high pass filter at 0.1 Hz with half amplitude cutoff at 0.05 Hz was applied to the continuous data to remove very low frequency activity not associated with neural activation without distorting the N400 component (Tanner et al., 2015). Next, epochs were extracted around Word 2, extending 2 s before and 4 s after word onset. Following linear detrending (described below), a zero-phase, non-causal low pass filter at 20 Hz with half amplitude cutoff at 22.5 Hz was applied to remove high frequency noise associated with movement such as eye blinks or jaw clenching. Next, the following processing steps were completed for each participant file separately (Supplementary Fig. 2):

1. Channels with low data quality were identified for interpolation using FASTER (Nolan et al., 2010) and pop_rejchan.

2. Low quality epochs were rejected using FASTER with a ± 3SD cutoff across three criteria, first computed separately by channel then averaged across channels: mean epoch amplitude difference from the respective channel's mean amplitude across epochs; absolute max–min voltage difference; and epoch voltage variance.
3. Initial independent components were computed using independent components analysis (ICA, runica, Makeig et al., 1995) to identify and remove components resembling blink and eye-movement artifacts.
4. The data was manually reviewed to identify any additional low-quality channels for interpolation or epochs for rejection. This additional interpolation and epoch rejection was applied to the data files created after step two.
5. An ICA was again run with both automatically and manually interpolated channels and rejected epochs removed, but without having removed any artifact components from the initial decomposition.
6. A final selection of blink and eye movement artifacts was subtracted.
7. Epochs were then pruned to include data from 500 ms before to 1000 ms after word onset.
8. A baseline mean amplitude was computed from −200 ms to 0 ms relative to word onset, separately for each channel and trial. This was then subtracted from the entire trial for each channel.
9. Next, each trial and channel were linearly detrended.
10. Finally, the data were baseline corrected as before.

2.6. Data analysis

Participant data were excluded for particular tasks within a visit if their accuracy in either condition was below 60%. In addition, participant data were excluded for a particular task if fewer than 15 epochs with correct responses remained in either condition following the offline processing described above. See Supplementary Fig. 1 for a CONSORT diagram specifying the frequency with which exclusion criteria were applied. In Visit 1, the number of control participants (n = 27) retained was 26 for lexical decision and 27 for the orthographic, semantic, and rhyme tasks. In Visit 2, 24 control participants were retained for all tasks. There was greater variability in the number of PWA participants (n = 26) retained across visits and tasks. In Visit 1, 25 PWA were retained for lexical decision and orthographic, 23 for semantic, and 20 for rhyme tasks. In Visit 2, 23 PWA were retained for lexical decision, 22 for orthographic and semantic, and 19 for rhyme tasks.

All dependent variables (accuracy, reaction time, N400 amplitude) were computed separately per participant, visit, task, and condition. For a given dependent variable, this amounted to eight averages per visit for a participant without any excluded data (four tasks * two conditions). Accuracy is reported as the proportion of correct responses out of 40 trials. Reaction time was computed from the average of all responses, regardless of accuracy. The N400 amplitude was calculated from Cz (located at the center midline of the head) as the mean amplitude from 350 ms to 550 ms after the second word in a pair was presented. The selected electrode location, latency (timing), and duration are typical for N400 studies (Šoškić et al., 2022). While the latency is towards the later end of the range, this is appropriate given the older mean age of our sample, since previous research has reported a slightly later N400 response in this group (Kutas & Iragui, 1998). Furthermore, among midline electrodes, this location and window maximized the N400 effect, averaged across groups, visits, and tasks (semantic and rhyme).

To analyze the N400 effect, or the differential response to matching vs. nonmatching stimuli, difference waves were calculated between the conditions in each task (e.g., for the semantic condition: congruent condition – incongruent condition = difference wave). This contrast (congruent – incongruent) was selected for interpretability (e.g., a more positive difference wave indicated a larger N400 effect), simplified comparisons across conditions, and downstream statistical modeling (especially for comparison with behavioral measures). After the difference waves were computed, mean amplitude at Cz was calculated as described above (referred to as the "semantic N400 effect" or "rhyme N400 effect" throughout the results). Finally, the difference waves for the semantic and rhyme tasks were combined, and mean amplitude at Cz was calculated for both tasks together (referred to as the "average N400 effect" hereafter).

2.6.1. RQ 1 – ERP responses and between group differences

To answer the first research question of whether the tasks elicited expected ERP responses and to examine between-group differences, a series of 2 (group) × 2 (condition) mixed effect ANOVAs was completed. For the lexical decision and orthographic tasks (analyzed separately since they served as control tasks), these ANOVAs were calculated for task accuracy, response time, and N400 amplitude. The semantic and rhyme tasks were analyzed together using 2 (group) × 2 (condition) × 2 (task) mixed effect ANOVAs for task accuracy, reaction time, and N400 amplitude. Prior to calculating ANOVAs, data were checked to ensure they met assumptions for normality and homoscedasticity. Greenhouse-Geiser corrected p-values are reported for this analysis to account for repeated measurement. T-tests were used to examine pairwise comparisons of statistically significant ANOVAs. Generalized eta-squared (η^2_G) was used to evaluate the effect-size of the ANOVA, while Hedge's *g* or Cohen's *d* were used to evaluate the effect-size of pairwise t-tests (Cohen's *d* was used for sample sizes over 20). For *g* and *d*, the lower bounds of effect sizes were defined as: small = 0.2; medium = 0.5; large = 0.8. For η^2_G, the lower bounds of effect sizes were defined as: small = 0.01; medium = 0.06; large = 0.14.

2.6.2. RQ 2 – reliability and stability of the N400

The second research question addressed the psychometric properties of the N400 response, specifically test–retest reliability and stability. Test-retest reliability and stability of N400 variables was calculated across participants for pairs of visits. Visits were separated by an average of 31.1 days ([min, max] = [25, 38]) for control participants and an average of 34.7 days [17, 62] for PWA; visit intervals were not significantly different between groups ($t(29.1) = 1.83$, $p = 0.077$). Test-retest reliability and stability were calculated separately for the respective conditions of the semantic (incongruent and congruent amplitude) and rhyme tasks (nonrhyme and rhyme amplitude), as well as the N400 effect within each task (semantic N400 effect, rhyme N400 effect). Finally, they were calculated for the average N400 effect, combining semantic and rhyme tasks. Results are also reported separately by participant group.

Test-retest reliability was evaluated using Pearson's correlations (*r*), which allowed us to determine the strength of the association between variables across two time points. Adequacy of test–retest reliability was defined as negligible ($r < 0.30$), low (*r* between 0.30 and 0.50), moderate (*r* between 0.50 and 0.70), high (*r* between 0.70 and 0.90), and very high ($r > 0.90$), as defined by Mukaka (2012).

Complementary to test–retest reliability, stability was estimated using intraclass correlation coefficients (ICCs; psych package, Revelle, 2022), which assumes the variables are repeated measures on the exact same scale and is therefore more sensitive to changes in magnitude (e.g., habituation to stimuli across repeated measures). The ICC was conducted specifying a two-way model with fixed 'raters' and absolute 'rater' agreement to identify the magnitude of the difference between performances at the two times. Adequacy of stability was defined as poor (ICC < 0.5), moderate (ICC between 0.5 and 0.75), good (ICC between 0.75 and 0.9), and excellent (ICC > 0.9), as suggested by Koo & Li (2016).

2.6.3. RQ 3 – relationship between N400 and cognitive-linguistic abilities in PWA

To address the third research question – examining the relationship

between N400 variables and cognitive-linguistic abilities – we conducted Spearman correlations as appropriate between behavioral test results and ERP variables. Based on the preceding reliability analysis, language-specific behaviors of interest (WAB-AQ, BNT, MC composite score, number of accurate/complete MCs, number of MC attempts, and number of errored MCs) were entered into the correlation analysis with condition-specific amplitudes (incongruent, nonrhyme). Correlation coefficients were interpreted according to Mukaka's standards described above (2012). All main concept scores were calculated as the sum of scores across five discourse tasks (two picture sequence descriptions, a picture scene description, a story retell, and a procedural explanation) from the Discourse Production Test (DPT).

3. RESULTS

3.1. RQ 1 – ERP responses and between group differences

3.1.1. Lexical decision task

To assess attentiveness and responsiveness in PWA relative to controls, we compared group accuracy (ACC) and reaction time (RT) during the lexical decision task using separate 2 (group) × 2 (condition) mixed effects ANOVAs. Results demonstrated significant main effects of condition (**ACC**: $F(1,49) = 21.6$, $p < 0.001$, $\eta^2_G = 0.20$; **RT**: $F(1,49) = 65.1$, $p < 0.001$, $\eta^2_G = 0.19$) and group (**ACC**: $F(1,49) = 6.66$, $p = 0.013$, $\eta^2_G = 0.06$; **RT**: $F(1,49) = 10.4$, $p = 0.002$, $\eta^2_G = 0.15$) such that accuracy was higher and reaction time was faster for words than nonwords for both groups. Additionally, healthy controls were more accurate and had faster reaction times than PWA. No significant interaction effects were observed. Finally, we compared amplitudes in the N400 window between the two conditions and found no significant main or interaction effects ($Fs < 1.8$, $ps > 0.187$, $\eta^2_G s < 0.03$). This aligned with our expectation that the lexical decision task would serve as a linguistic control task which did not elicit an N400 effect. Please see Supplementary section "Results – Lexical Decision Task" and Supplementary Fig. 3 for a complete reporting of these analyses and results.

3.1.2. Orthographic task

Behavioral results from the orthographic task mirrored that of the lexical decision task, again demonstrating that participants were engaged and successful at completing the task. Similarly, a 2 (group) × 2 (condition) mixed effects ANOVA examining N400 amplitudes supported our hypothesis that this task would not elicit an N400 effect, with

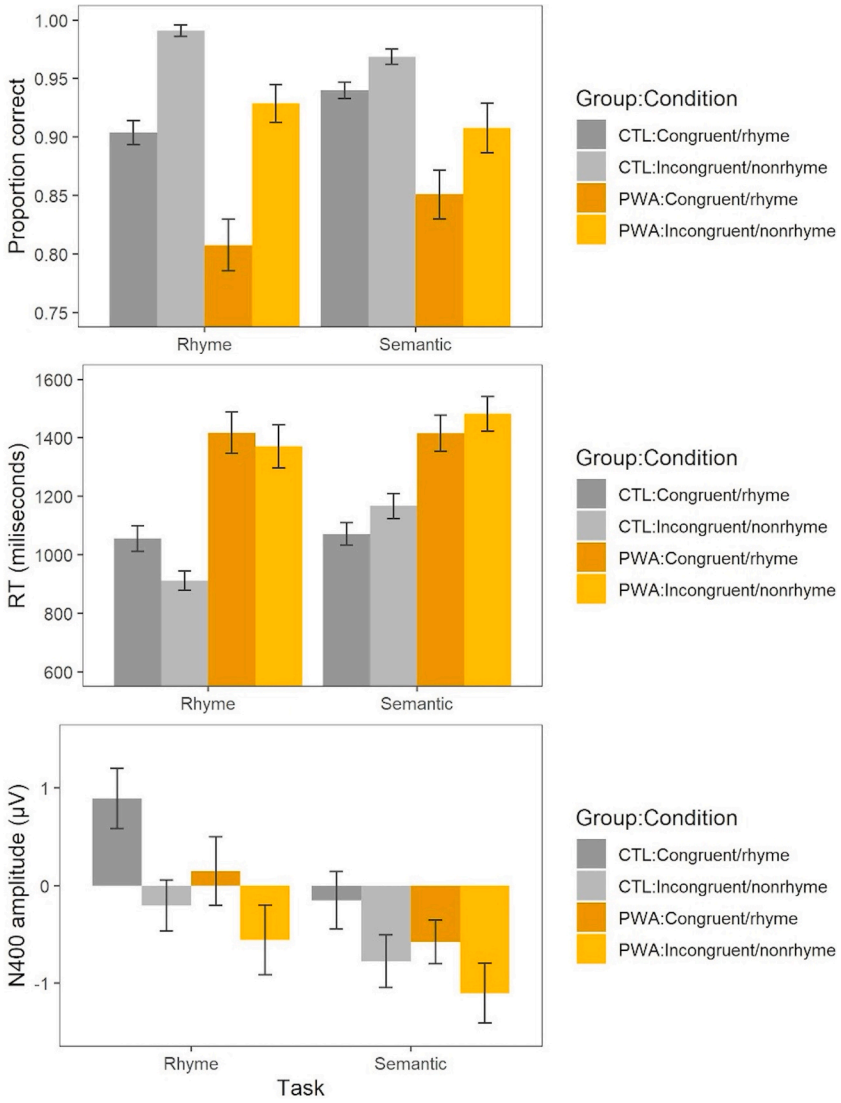


Fig. 1. Mean response accuracy (top), reaction time (RT) (middle), and ERP amplitude at Cz from 350 ms to 550 ms (bottom) for paired-word phonological and semantic comparison tasks. The more saturated colors represent the matching condition regardless of task. Error bars represent ± 1 standard error of the mean. CTL–healthy controls. PWA–persons with aphasia.

no significant main or interaction effects observed ($Fs < 2$, $ps > 0.164$, $\eta^2_G s < 0.04$). Please see Supplementary section "Results – Orthographic Task" and Supplementary Fig. 4 for a complete reporting of these analyses and results.

3.1.3. Semantic and rhyme tasks

The next analyses focused on the semantic and rhyme tasks. ANOVAs (described below) were conducted which included only participants with acceptable data on both tasks. A 2 (group) × 2 (condition) × 2 (task) mixed effects ANOVA was conducted on accuracy (Fig. 1).This revealed significant main effects of group ($F(1,44) = 30.5$, $p < 0.001$, $\eta^2_G = 0.23$) and condition ($F(1,44) = 63.0$, $p < 0.001$, $\eta^2_G = 0.20$) but no significant effect of task ($F(1,44) = 2.58$, $p = 0.115$, $\eta^2_G = 0.01$). The only significant interaction was between condition and task ($F(1,44) = 15.2$, $p < 0.001$, $\eta^2_G = 0.06$; all other two- and three-way interactions were not significant: $Fs(1,44) < 0.01$, $ps > 0.929$, $\eta^2_G s < 0.01$). Pairwise comparisons revealed that PWA performed worse than control participants on both tasks (**semantic**: $t(22.3) = -3.7$, $p = 0.001$, $g = -1.2$; **rhyme**: $t(22.5) = -4.4$, $p < 0.001$, $g = -1.5$). Congruent and rhyme conditions yielded lower performance across both tasks and both groups, compared to incongruent and nonrhyme, respectively. Across both groups, performance was worse for congruent than incongruent conditions of the semantic task (**PWA**: $t(22) = -2.1$, $p = 0.05$, $d = -0.43$; **CTL**: $t(26) = -3.8$, $p = 0.001$, $d = -0.72$). However, the rhyme condition was disproportionately hard compared to the nonrhyme condition of the rhyme task (**PWA**: $t(19) = -5.7$, $p < 0.001$, $d = -1.3$; **CTL**: $t(26) = -7.7$, $p < 0.001$, $d = -1.5$).

A 2 (group) × 2 (condition) × 2 (task) mixed effects ANOVA was conducted on reaction time (Fig. 1). This revealed significant main effects of group ($F(1,44) = 35.3$, $p < 0.001$, $\eta^2_G = 0.36$) and task ($F(1,44) = 6.92$, $p = 0.012$, $\eta^2_G = 0.02$) but not condition ($F(1,44) = 5.49$, $p = 0.023$, $\eta^2_G < 0.01$). The condition by task interaction was significant ($F(1,44) = 28.3$, $p < 0.001$, $\eta^2_G = 0.03$) as was the group by condition by task interaction ($F(1,44) = 6.92$, $p = 0.012$, $\eta^2_G = 0.01$). Notwithstanding the three-way interaction, PWA were slower for both tasks (**semantic**: $t(29.2) = 4.40$, $p < 0.001$, $g = 1.4$; **rhyme**: $t(26.4) = 5.86$, $p < 0.001$, $g = 1.9$). The condition effect was reversed in the semantic compared to rhyme task, and this reversal was more pronounced for control participants. For control participants, RT for incongruent trials was slower ($t(26) = -2.5$, $p = 0.019$, $d = -0.48$) whereas RT in nonrhyme trials was faster ($t(26) = 4.2$, $p < 0.001$, $d = 0.80$. PWA showed nonsignificant trends in the same directions (**semantic**: $t(22) = -1.46$, $p = 0.159$, $d = -0.30$; **rhyme**: $t(19) = 0.977$, $p = 0.341$, $d = 0.22$).

3.1.4. N400 component

A 2 (group) × 2 (condition) × 2 (task) mixed effects ANOVA was conducted on N400 amplitudes (Fig. 1). This revealed significant main effects of condition ($F(1,44) = 39.5$, $p < 0.001$, $\eta^2_G = 0.06$) and task ($F(1,44) = 19.0$, $p < 0.001$, $\eta^2_G = 0.04$) but not group ($F(1,44) = 0.93$, $p = 0.340$, $\eta^2_G = 0.02$). None of the two- or three-way interactions were significant ($Fs(1,44) < 2.35$, $p > 0.13$, $\eta^2_G s < 0.01$). The main effect of task was driven by more negative amplitudes in the semantic task than the rhyme task.

PWA and control participants showed N400 effects (congruent < incongruent and rhyme < nonrhyme negative amplitudes) on the semantic (**PWA**: $t(22) = 2.6$, $p = 0.018$, $d = 0.54$; **CTL**: $t(26) = 3.5$, $p = 0.002$, $d = 0.68$) and rhyme tasks (**PWA**: $t(19) = 2.7$, $p = 0.015$, $d = 0.60$; **CTL**: $t(26) = 6.0$, $p < 0.001$, $d = 1.15$), respectively. However, no significant differences between PWA and control participants were observed for N400 effects on the semantic or rhyme tasks ($ps > 0.108$; Supplementary Fig. 5). Finally, the average of semantic and rhyme N400 effects was calculated by group. Consistent with the nonsignificant interactions in the ANOVA, control participants did not have a significantly larger task-averaged N400 effect ($t(40.3) = -1.03$, $p = 0.311$, $g = -0.30$) than PWA.

3.2. RQ 2 – reliability and stability of the N400

3.2.1. Test-retest reliability

Table 2 and Fig. 2 present Pearson's correlations of N400 amplitudes across visits 1 and 2 separately by participant group and task to evaluate test–retest reliability. Note that participants were removed listwise if their data were excluded or missing for either visit, separately by task. All condition-specific N400 amplitudes were significantly positively correlated for the semantic (*congruent* **PWA**: $r = 0.46$, $p < 0.05$; **CTL**: $r = 0.68$, $p < 0.001$; *incongruent* **PWA**: $r = 0.73$, $p < 0.001$; **CTL**: $r = 0.78$, $p < 0.001$) and rhyme (*rhyme* **PWA**: $r = 0.66$, $p < 0.01$; **CTL**: $r = 0.9$, $p < 0.001$; *nonrhyme* **PWA**: $r = 0.82$, $p < 0.001$; **CTL**: $r = 0.71$, $p < 0.001$) tasks for both groups with respect to test–retest reliability. For healthy controls, test–retest reliability ranged from moderate (congruent) to high (incongruent and nonrhyme) to very high (rhyme). For PWA, test–retest reliability ranged from low (congruent), to moderate (rhyme) to high (incongruent, nonrhyme).

In contrast to the condition specific correlations, test–retest reliability for the N400 task effect amplitudes were poorer for both groups. Moderate positive correlations were observed for healthy controls for the rhyme N400 effect ($r = 0.58$, $p < 0.01$) and the task averaged N400 effect ($r = 0.66$, $p < 0.001$). For PWA, low positive (and non-significant) correlations were observed for the rhyme ($r = 0.35$, $p > 0.05$) and task-averaged N400 effect ($r = 0.45$, $p > 0.05$). For both groups, the semantic N400 effect was not significantly correlated between the two time points (**PWA**: $r = 0.04$, $p > 0.05$; **CTL**: $r = 0.34$, $p > 0.05$).

3.2.2. Stability

ICC correlations of N400 amplitudes across visits 1 and 2 are presented in Table 2, separately by group and task to evaluate stability. As above, participants were removed listwise if data were excluded or missing. For the condition specific amplitude analyses, significant positive correlations were observed for both conditions in the semantic (*congruent* **PWA**: ICC = 0.44, $p < 0.05$; **CTL**: ICC = 0.67, $p < 0.001$; *incongruent* **PWA**: ICC = 0.73, $p < 0.001$; **CTL**: ICC = 0.78, $p < 0.001$) and rhyme (*rhyme* **PWA**: ICC = 0.62, $p < 0.01$; **CTL**: ICC = 0.9, $p < 0.001$; *nonrhyme* **PWA**: ICC = 0.82, $p < 0.001$; **CTL**: ICC = 0.71, $p < 0.001$) tasks for both groups. Controls demonstrated excellent stability in the rhyme condition, good stability in the incongruent condition, and moderate stability in the congruent and nonrhyme conditions. PWA demonstrated good stability in the nonrhyme condition, moderate stability in the incongruent and rhyme conditions, and poor stability in the congruent condition.

Similar to the reliability results, the N400 task effect amplitudes also had lower stability than the condition specific amplitudes. Healthy controls demonstrated moderate stability for the rhyme N400 (ICC = 0.58, $p < 0.01$) and task-averaged N400 effect (ICC = 0.64, $p < 0.001$), but poor (and non-significant) stability for the semantic N400 effect (ICC = 0.33, $p > 0.05$). PWA demonstrated poor stability for the task-averaged N400 effect (ICC = 0.45, $p < 0.05$) and poor (non-significant) stability for the semantic (ICC = 0.04, $p > 0.05$) and rhyme N400 (ICC = 0.35, $p > 0.05$) effects.

3.3. RQ 3 – relationship between N400 and cognitive-linguistic abilities in PWA

Because condition specific amplitudes are more reliable than difference scores (e.g., N400 effects), and given our finding that the test–retest reliability and stability of condition-specific N400 amplitudes were higher than those of the N400 effects (Fig. 2, Table 2), we conducted correlations between condition-specific N400 amplitudes and linguistic measures for PWA. We focused on the N400 amplitudes from the incongruent and nonrhyme conditions (from semantic and rhyme tasks, respectively) as these both had moderate reliability for PWA and controls, as well as higher reliability among PWA compared to the opposite conditions (congruent and rhyme, respectively). Results

Table 2
Pearson's *r* and intraclass correlation coefficients between N400 ERP amplitudes at Visit 1 and Visit 2 for control participants and persons with aphasia, for the Semantic and Rhyme tasks.

Reliability statistic	Group	ERP Source						
		Incon-gruent	Con-gruent	Semantic effect	Non-rhyme	Rhyme	Rhyme effect	Task Average Effect
r	CTL	0.78*** [0.55, 0.9]	0.68*** [0.38, 0.85]	0.34 [−0.08, 0.65]	0.71*** [0.43, 0.87]	0.9*** [0.77, 0.95]	0.58** [0.23, 0.8]	0.66*** [0.35, 0.84]
r	PWA	0.73*** [0.42, 0.88]	0.46* [0.02, 0.75]	0.04 [−0.41, 0.47]	0.82*** [0.56, 0.93]	0.66** [0.28, 0.86]	0.35 [−0.14, 0.7]	0.45 [−0.05, 0.78]
ICC	CTL	0.78*** [0.55, 0.9]	0.67*** [0.37, 0.84]	0.33 [−0.07, 0.65]	0.71*** [0.43, 0.86]	0.9*** [0.78, 0.95]	0.58** [0.23, 0.79]	0.64*** [0.32, 0.82]
ICC	PWA	0.73*** [0.43, 0.88]	0.44* [0.01, 0.74]	0.04 [−0.4, 0.46]	0.82*** [0.57, 0.93]	0.62** [0.23, 0.84]	0.35 [−0.13, 0.69]	0.45* [−0.04, 0.76]

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.
Brackets contain the lower and upper bounds of the 95% confidence intervals of each respective correlation coefficient. Note the respective sample sizes for the separate tasks and groups: Controls, $N = 24$ for Semantic and Rhyme; Persons with Aphasia: $N = 20$ for Semantic and $N = 18$ for Rhyme. CTL: healthy controls; PWA: persons with aphasia.

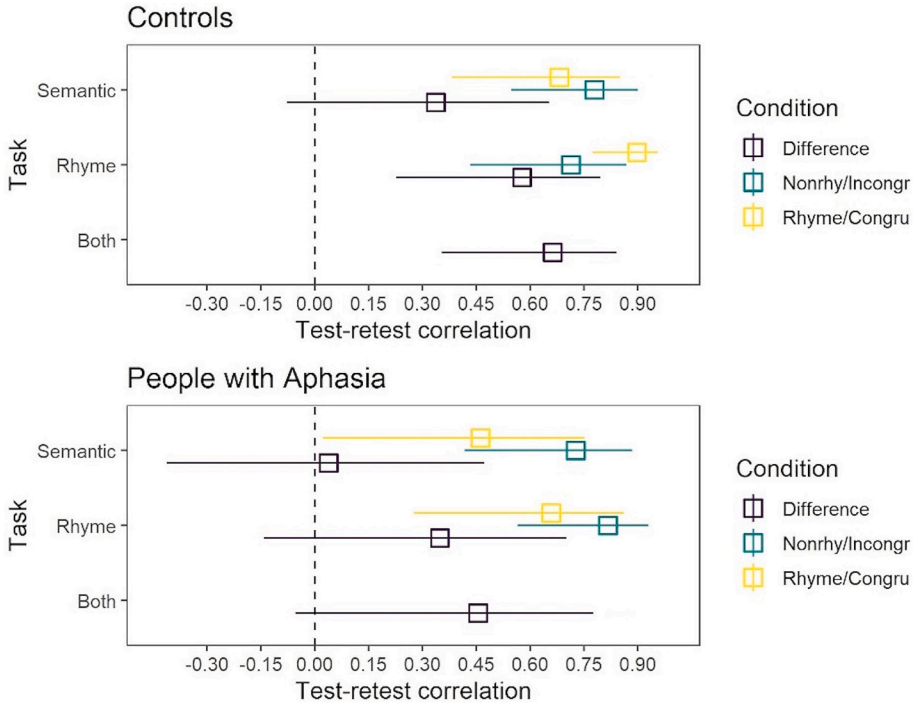


Fig. 2. Test-retest correlation between visits 1 and 2 of N400 ERP amplitudes (see methods for details) for the paired-word phonological and semantic comparison tasks. Correlations were repeated for task conditions (match–yellow; or nonmatch–blue) as well as task effects–purple, for both tasks. Correlation was also repeated for the overall task effect averaged across tasks (Both). Error bars represent lower and upper bounds of the 95% confidence interval for Pearson's *r* coefficient.

demonstrated that there was a small positive correlation ($r = 0.48$, $p < 0.05$) between N400 amplitude during the incongruent condition and the MC composite score. In addition, the incongruent N400 amplitude was moderately positively correlated ($r = 0.51$, $p < 0.05$) with the number of main concepts attempted (Supplementary Table 2; Fig. 3). No significant correlations were observed between ERP measures and WAB-R AQ, BNT, number of AC main concepts, and number of errored main concepts. Additional analyses revealed that N400 effects from rhyme and semantic tasks were not significantly correlated to any of the cognitive-linguistic measures (Supplementary Table 2).

3.4. Summary of results

Across tasks, both PWA and healthy controls demonstrated high task engagement, with controls showing faster and more accurate responses overall. As expected, the lexical decision and orthographic control tasks did not elicit N400 effects in either group, supporting their role as validated non-N400 control paradigms. In contrast, both the semantic and rhyme tasks elicited robust N400 responses in PWA and controls, with significant effects of condition and task but no significant group differences in N400 amplitude or effect size. Behavioral performance differed across tasks and conditions, with greater difficulty observed for rhyme judgments relative to semantic judgments, particularly in accuracy, and task-specific reversals in reaction time patterns across conditions. Test-retest analyses showed that condition-specific N400 amplitudes demonstrated moderate to high reliability and stability across visits in both groups, whereas N400 effect (difference score) measures were less reliable, particularly for PWA. Finally, in PWA, condition-specific N400 amplitudes from the semantic incongruent condition were significantly correlated with discourse-based language measures, whereas no significant relationships were observed with standardized aphasia test scores.

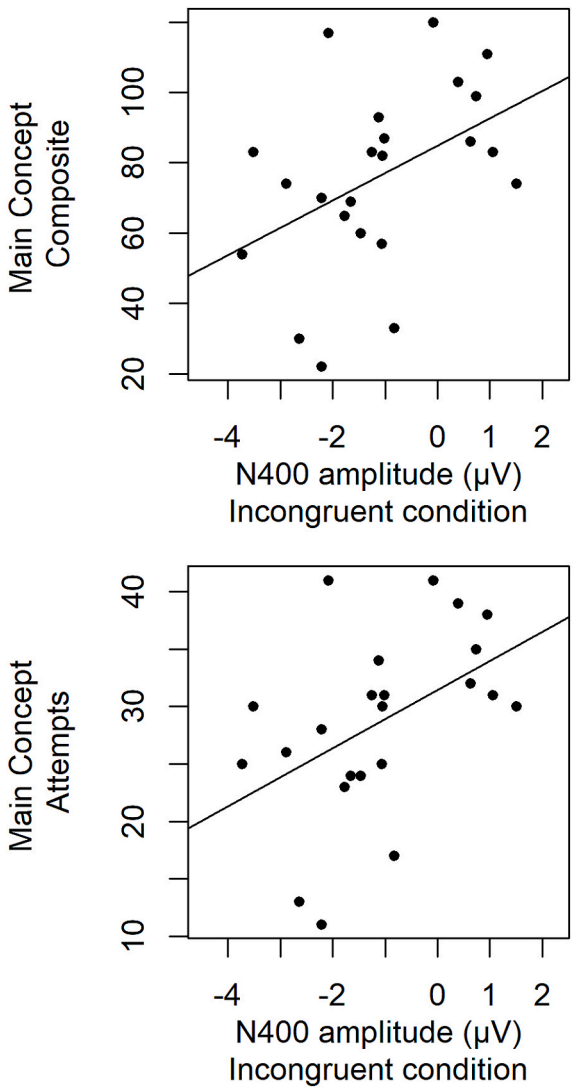


Fig. 3. Scatterplots showing the relationship between N400 amplitude in the incongruent condition and main concept composite scores (top) and main concept attempts (bottom).

4. Discussion

4.1. Validation of word-level semantic and rhyme judgement task to elicit the N400

We developed two single-word linguistic control tasks and two single-word linguistic N400 elicitation tasks for use with PWA and healthy controls. While all four tasks were able to be completed by both healthy controls and the majority of PWA, controls exhibited faster and more accurate behavioral responses across all tasks. Despite these differences, most PWA also successfully completed the behavioral tasks, albeit with greater variability in reaction time and accuracy. As hypothesized, neither our lexical decision nor our orthographic task elicited an N400 response. While some previous studies using lexical decision tasks in PWA have reported an N400 response, those studies included priming or other manipulations to produce the response (e.g., Swick, 1998; Swick & Knight, 1999). Our findings are consistent with Pulvermüller and colleagues (2004), who used a lexical decision task without manipulations and did not observe an N400 response. Importantly, PWA performed similarly to healthy controls on both the lexical decision and orthographic tasks, suggesting that PWA process stimuli in expected, obligatory ways, without atypical responses to semantic or phonological information, despite their brain injury. These findings lay the groundwork for further exploration of word-level semantic and phonological N400 effects in PWA.

Validated control tasks are critical for methodologically rigorous ERP research, to ensure that targeted neural components are elicited only in the expected or intended contexts. Without prior validation, findings that a task does not elicit an N400 have limited interpretability since null findings could be related to issues with task design (e.g., poor stimulus selection, unclear instructions, timing issues, etc.) rather than *real* underlying processing differences. Moreover, validated control tasks offer valuable contrasts during data analysis. For example, the orthographic task could serve as a baseline comparison for semantic and rhyme tasks, effectively removing activation linked to basic visual processing and clarifying activation specific to the N400 manipulation. Similarly, the lexical decision task could help control for general linguistic processing, enabling clearer identification of N400-specific variables.

4.1.1. Differences and similarities between the N400 in healthy controls and PWA

Both healthy controls and PWA produced a clear N400 response and an N400 effect during the semantic and rhyme tasks. There were no significant differences between the two groups on the N400 amplitudes or the magnitude of the N400 effect for either task, or when the tasks were averaged together. This is evident in the ERP traces and topographic plots (Fig. 4). Our findings are consistent with previous research (e.g., Räling, 2016; Khachatryan et al., 2018) reporting null results for the comparison of N400 amplitudes and/or effects between controls and PWA for single-word semantic tasks. However, Robson and colleagues (2017) did report significant differences between controls and individuals with Wernicke's aphasia on the magnitude of N400 effect. This discrepancy may be due to differences in the degree of comprehension impairment in our study as compared to Robson et al. (2017), since N400 amplitude is correlated with comprehension (e.g., D'Arcy et al., 2003). Participants whose data were able to be analyzed in the current study were primarily diagnosed with aphasia types not associated with significant comprehension deficits. However, further research with larger sample sizes to facilitate comparisons across subtypes is needed to validate this hypothesis.

Together, our findings and prior studies indicate that semantic and phonological processing mechanisms remain at least partially intact in PWA, even after brain injury, and that the N400 provides direct neural evidence of language processing abilities that may not be reflected in behavior (e.g., naming, comprehension, discourse). Further, these results contribute to continued refinement for models of language functioning and recovery in PWA. While traditional models still often frame aphasia primarily as a disruption of specific linguistic systems (e.g., semantics, syntax, phonology, etc.), this view is incomplete. The preservation of N400 responses in PWA supports network-based or resource-limitation models, which suggest that core language systems remain intact (or partially intact) but are less efficiently accessed or coordinated.

Interestingly, both groups completed the nonrhyme condition with the highest accuracy, and the rhyme condition with the lowest accuracy, suggesting that the latter task is particularly challenging for both groups. In addition, both groups showed a reversal in reaction time performance across tasks, such that in the semantic task, reaction times were faster to the congruent (matching) condition, while in the rhyme tasks, reaction times were faster to the incongruent (nonrhyme) condition. While it is possible that these results are driven in part by the stimuli (since it is challenging to identify nonrhyming words that appear to rhyme, and vice versa), the N400 effects observed in response to the task suggest that this is not the only factor driving performance. Indeed, while semantic processing and judgment is a routine aspect of everyday life, there are far fewer occasions when we consciously process and judge phonological features in everyday life. This likely also explains the

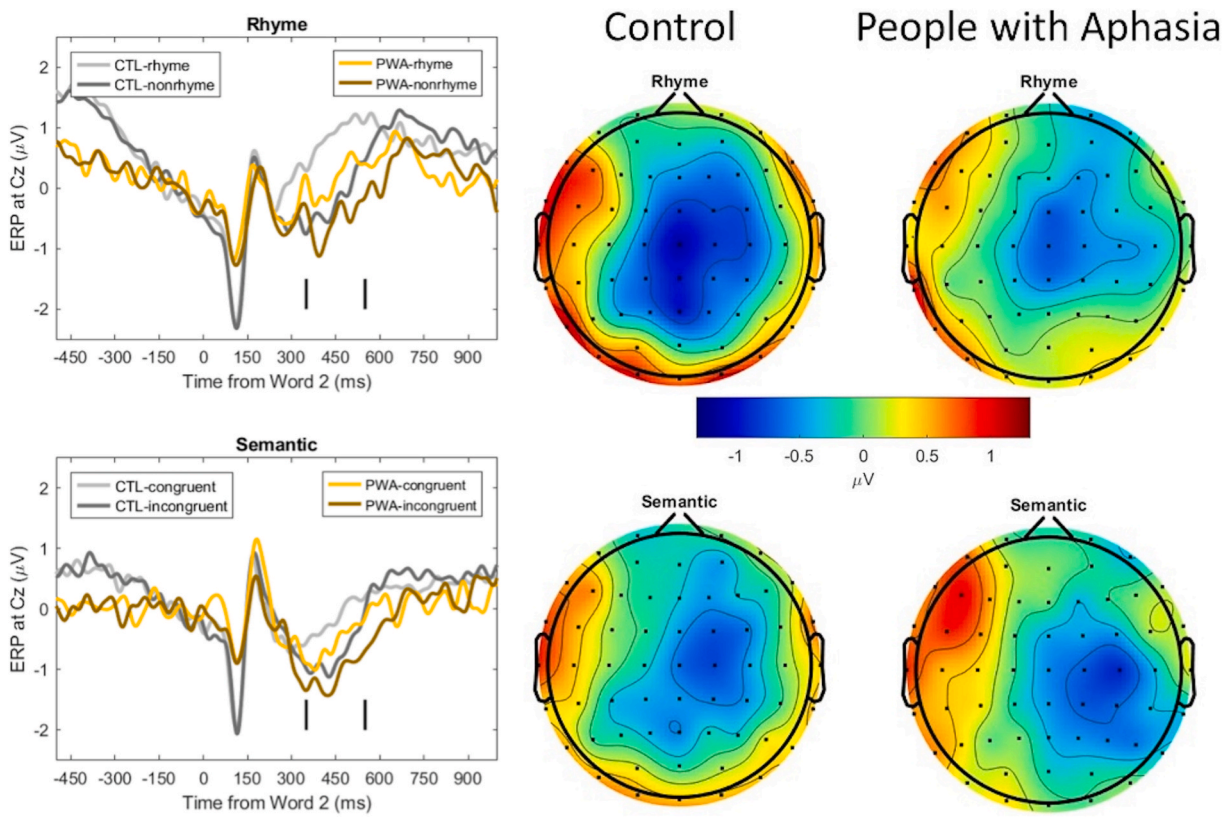


Fig. 4. Mean ERP time courses at Cz for Word 2 from the paired-word phonological (upper) and semantic (lower) comparison tasks (left) and topographic maps of N400 effects (non-match minus match) from 350 ms to 550 ms (right). Vertical bars at 350 ms and 550 ms represent the window for quantifying N400 amplitudes. CTL–healthy controls. PWA–persons with aphasia.

relatively more robust N400 effect observed for both groups in the rhyme task. These qualitatively contrasting results further support the notion that the semantic and rhyme tasks are tapping into different systems and can be used to evaluate different aspects of linguistic processing.

4.2. *Reliability and stability of the N400 in healthy controls and PWA*

Our results are broadly consistent with prior studies on the test–retest reliability of N400 amplitudes measured from a word-pair semantic congruency task among healthy controls (Kiang, et al., 2013) and people with schizophrenia (Boyd, et al., 2014). In general, condition-specific amplitudes have greater reliability than task effects, otherwise known as difference or delta scores (Imburgio, et al., 2020), and N400 variables have higher reliability among controls than patient groups. The rhyme task appeared to elicit more reliable N400 amplitudes and effects. However, the correlation of the average task effect was markedly enhanced relative to either semantic or rhyme task effect alone. Across groups, there appears to be no pattern as to whether the congruent conditions elicited a more reliable N400 amplitude compared to the incongruent conditions. Our results suggest that these N400 tasks may be appropriate to use over time, but that care must be taken when selecting the variables for investigation. Specifically, for both controls and PWA, the incongruent and nonrhyme conditions had high reliability appropriate for repeated measures. The rhyme condition had very high test–retest reliability for controls, but only moderate test–retest reliability for PWA. This suggests that while the rhyme condition may be suitable, it warrants closer scrutiny, particularly when comparing results. If the outcomes from the rhyme condition differ significantly from those of the incongruent and nonrhyme conditions, such differences should be interpreted with caution.

Stability was also higher for controls than PWA. Using the 95% confidence interval ranges of the ICCs, stability was moderate-good for the incongruent condition and good–excellent for the rhyme condition for controls. In contrast, PWA demonstrated acceptable (moderate-good) stability in the nonrhyme condition only. On the incongruent condition, PWA demonstrated borderline stability ranging from poor to good (0.43–0.88). Taken together, these results suggest that the nonrhyme condition is the most appropriate option for repeated measurement with PWA, while the rhyme and incongruent conditions may be most appropriate for repeated measurement with healthy controls. Note, however, this does not mean a task should be entirely or primarily composed of nonrhyming or incongruent word pairs – the contrasting conditions are essential for a robust N400 response.

4.3. *Relationship between the N400 and language measures in PWA*

We hypothesized that discourse measures, which are more directly related to real world communication than traditional, standardized aphasia assessments, would be related to language processing as indexed by N400 tasks. Consistent with this prediction, we observed significant correlations between discourse measures and condition-specific N400 amplitudes in PWA. This is a key finding, as it further supports the potential utility of the N400 as a biomarker for everyday communicative functioning in PWA. To our knowledge, this is the first study to relate N400 to discourse production abilities in PWA.

Both the main concept (MC) composite score, an indicator of the overall informativeness (or gist conveyed) of a discourse sample, and MC attempts, which indexes how many MCs an individual attempted to produce, regardless of accuracy or completeness, were positively correlated with the amplitude during the incongruent condition of the semantic task. From a theoretical perspective, this correlation was expected, since both the MC composite and MC attempt scores strongly index lexical-semantic access. In contrast, neither MC score is strongly

impacted by phonological processing, since individuals can receive credit for producing words with phonological errors, if they are intelligible in context. This is consistent with the lack of significant correlations between the nonrhyme condition and discourse measures.

These results are also consistent with previous research investigating the relationship between discourse production and spectral resting-state EEG (Dalton et al., 2021). Specifically, discourse informativeness in PWA, as measured by the MC composite score, was positively correlated with power in high-frequency beta and alpha bands, and negatively correlated with low-frequency theta band. Here, we extend the previous findings beyond resting-state EEG, to task-specific ERPs. This link between the N400 and real-world functional communication abilities in PWA is an important step in bridging the gap between neurophysiological processes and everyday communication challenges. Further examination of the N400 in relation to discourse tasks in PWA may help us to better understand cognitive mechanisms underlying real-time language processing, such as prediction, integration, and ambiguity resolution.

The fact that discourse variables correlated with the N400 while general language or naming scores did not, further highlights the importance of prioritizing discourse measures as a primary outcome in aphasia research. Discourse performance offers a more precise reflection of the functional communication outcomes that PWA value most, and the presence of a neurophysiological marker linked to this real-world behavior would provide an objective, quantifiable measure that can enhance both assessment and treatment strategies. This is a promising finding in our relatively small sample, and this relationship should be explored further and replicated in larger samples to strengthen its theoretical and clinical implications.

4.4. Limitations and future directions

Our results support the continued relevance of the N400 while highlighting some avenues for further research. First, while most PWA were able to complete the tasks, a few PWA struggled with task demands, and several were excluded from the analyses, either because of low accuracy or response patterns that suggested they were unsure of how to complete the task. This highlights the need for further development of paradigms, tasks, or analytical approaches that can accommodate the full spectrum of PWA type and severity, including those who produce fewer accurate responses. Using a paradigm which does not require a button-press response might have the dual benefits of being easier for all individuals with aphasia to complete while reducing contralateral hemisphere signals associated with motor activation. Additionally, investigating the use of auditory or auditory and visual paradigms may support task completion for individuals with more severe aphasia. We suspect that the relationship between the N400 and language measures might have been even stronger if the entire continuum of individuals had been included in the analysis, offering a more complete understanding of how aphasia severity relates to N400 response.

A related limitation concerns sample size and participant heterogeneity. As is typical in clinical electrophysiological research, the present sample includes substantial variability in lesion location, aphasia subtype, and behavioral profiles. While this limits subtype-specific inferences, robust N400 effects were nonetheless observed at the group level, and the relationship between N400 and cognitive-linguistic measures suggests sensitivity to meaningful individual differences rather than effects driven by a small subset of participants. These findings should thus be considered an initial characterization of N400 responses across a heterogeneous aphasia sample, supporting their reliability and functional relevance while also highlighting the importance of future studies with larger samples that will permit stratification by subtype, severity, and/or lesion characteristics.

Of note, we observed large N100-P200 complexes in our data (see Fig. 4, 100–200 ms). Significant group differences in N100 were observed, but no significant task or condition effects, or interactions, were observed (see Supplementary Materials Results – N100 ERPs and Supplementary Fig. 6). While not relevant in the context of the current investigation, it may be of interest in the future to examine whether this ERP complex is informative with respect to language processing in controls and/or PWA or whether it functions as a purely sensory marker of task-related processing.

5. Conclusions

This study makes significant contributions to N400 research by validating new linguistic tasks for both PWA and healthy controls. By measuring N400 responses in both semantic and phonological (rhyme) contexts, the study offers critical insights into how PWA process linguistic stimuli compared to healthy controls, revealing that while both groups exhibit typical N400 effects, PWA show greater variability in performance. We also characterized performance on tasks that can serve as validated controls for future N400 studies (i.e., lexical decision, orthographic), enabling more reliable interpretations of N400 effects by distinguishing between basic visual and linguistic processing. We assessed the test–retest reliability and stability of N400 amplitudes, with findings that highlight the importance of carefully selecting conditions when using N400 measures in repeated assessments of PWA. This emphasizes both the potential benefits and challenges of using these measures in longitudinal aphasia studies. Moreover, we showed that N400 responses correlated with discourse production variables in PWA, which are more closely aligned with real-world communication outcomes than traditional aphasia assessments or impairment-based metrics more commonly used in EEG research. These results also prompt important questions about how N400 responses may vary across the continuum of aphasia severity, suggesting that future studies could benefit from including a broader range of participants to better understand the relationship between neural responses and language abilities in PWA. Overall, this work represents a crucial step in bridging neurophysiological processes with functional language use in aphasia.

Funding Sources

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number P20GM109089 and the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award number R01DC018282-01A1.

CRediT authorship contribution statement

Sarah Grace Dalton: Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Mark Lavelle:** Writing – original draft, Visualization, Validation, Formal analysis. **James F. Cavanagh:** Writing – review & editing, Resources, Methodology, Formal analysis. **Jessica D. Richardson:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jessica Richardson and James Cavanagh reports financial support was provided by NIH National Institute of General Medical Sciences. Jessica Richardson reports financial support was provided by National Institute on Deafness and Other Communication Disorders. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bandl.2026.105752.

Data availability

Data will be made available on request.

References

Aerts, A., Batens, K., Santens, P., Van Mierlo, P., Huysman, E., Hartsuiker, R., Hemelsoet, D., Duyck, W., Raedt, R., Van Roost, D., & De Letter, M. (2015). Aphasia therapy early after stroke: Behavioural and neurophysiological changes in the acute and post-acute phases. *Aphasiology, 29*(7), 845–871. https://doi.org/10.1080/02687038.2014.996520

Arheix-Parras, S., Glize, B., Guehl, D., & Python, G. (2023). Electrophysiological changes in patients with post-stroke aphasia: A systematic review. *Brain Topography, 36*(2), 135–171. https://doi.org/10.1007/s10548-023-00941-4

Barbancho, M. A., Berthier, M. L., Navas-Sánchez, P., Dávila, G., Green-Heredia, C., García-Alberca, J. M., Ruiz-Cruces, R., López-González, M. V., Dawid-Milner, M. S., Pulvermüller, F., & Lara, J. P. (2015). Bilateral brain reorganization with memantine and constraint-induced aphasia therapy in chronic post-stroke aphasia: An ERP study. *Brain and Language, 145*, 1–10. https://doi.org/10.1016/j.bandl.2015.04.003

Barwood, C. H., Murdoch, B. E., Whelan, B. M., Lloyd, D., Riek, S., O'Sullivan, J. D., Coulthard, A., & Wong, A. (2011). Modulation of N400 in chronic non-fluent aphasia using low frequency repetitive transcranial magnetic stimulation (rTMS). *Brain and Language, 116*(3), 125–135. https://doi.org/10.1016/j.bandl.2010.07.004

Barwood, C. H., Murdoch, B. E., Whelan, B. M., O'Sullivan, J. D., Wong, A., Lloyd, D., Riek, S., & Coulthard, A. (2012). Longitudinal modulation of N400 in chronic non-fluent aphasia using low-frequency rTMS: A randomised placebo controlled trial. *Aphasiology, 26*(1), 103–124. https://doi.org/10.1080/02687038.2011.617812

Boyd, J. E., Patriciu, I., McKinnon, M. C., & Kiang, M. (2014). Test–retest reliability of N400 event-related brain potential measures in a word-pair semantic priming paradigm in patients with schizophrenia. *Schizophrenia Research, 158*(1–3), 195–203. https://doi.org/10.1016/j.schres.2014.06.018

Brookshire, R. H., & Nicholas, L. E. (1997). *Discourse comprehension test: Test manual*. BRK Publishers.

Chatrian, G. E., Lettich, E., & Nelson, P. L. (1985). Ten percent electrode system for topographic studies of spontaneous and evoked EEG activities. *American Journal of EEG Technology, 25*(2), 83–92. https://doi.org/10.1080/00029238.1985.11080163

Cocquyt, E. M., Van Laeken, H., van Mierlo, P., & De Letter, M. (2023). Test–retest reliability of electroencephalographic and magnetoencephalographic measures elicited during language tasks: A literature review. *European Journal of Neuroscience, 57*(8), 1353–1367. https://doi.org/10.1111/ejn.15948

Dalton, S. G., Cavanagh, J. F., & Richardson, J. D. (2021). Spectral resting-state EEG (rsEEG) in chronic aphasia is reliable, sensitive, and correlates with functional behavior. *Frontiers in Human Neuroscience, 15*, Article 624660. https://doi.org/10.3389/fnhum.2021.624660

Dalton, S. G., Hubbard, H. I., & Richardson, J. D. (2020). Moving toward non-transcription based discourse analysis in stable and progressive aphasia. *Seminars in Speech and Language, 41*, 032–044. https://doi.org/ 10.1055/s-0039-3400990.

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

D'Arcy, R. C., Marchand, Y., Eskes, G. A., Harrison, E. R., Phillips, S. J., Major, A., & Connolly, J. F. (2003). Electrophysiological assessment of language function following stroke. *Clinical Neurophysiology, 114*(4), 662–672. https://doi.org/10.1016/S1388-2457(03)00007-5

Davies, M. (2008) *The Corpus of Contemporary American English (COCA)*. Available online at https://www.english-corpora.org/coca/.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Finnigan, S., Wong, A., & Read, S. (2016). Defining abnormal slow EEG activity in acute ischaemic stroke: Delta/alpha ratio as an optimal QEEG index. *Clinical Neurophysiology, 127*(2), 1452–1459. https://doi.org/10.1016/j.clinph.2015.07.014

Fromm, D., Forbes, M., Holland, A., & Macwhinney, B. (2020). Using AphasiaBank for discourse assessment. *Seminars in Speech and Language, 41*, 010–019. https://doi.org/ 10.1055/s-0039-3399499.

Gorišek, V. R., Isoski, V. Z., Belič, A., Manouilidou, C., Koritnik, B., Bon, J., Meglič, N. P., Vrabec, M., Žibert, J., Repovš, G., & Zidar, J. (2016). Beyond aphasia: Altered EEG connectivity in Broca's patients during working memory task. *Brain and Language, 163*, 10–21. https://doi.org/10.1016/j.bandl.2016.08.003

Imburgio, M. J., Banica, I., Hill, K. E., Weinberg, A., Foti, D., & MacNamara, A. (2020). Establishing norms for error-related brain activity during the arrow Flanker task among young adults. *NeuroImage, 213*, Article 116694. https://doi.org/10.1016/j.neuroimage.2020.116694

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston Naming Test*. Pro-ed.

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

Khachatryan, E., Wittevrongel, B., De Keyser, K., De Letter, M., & Hulle, M. M. V. (2018). Event related potential study of language interaction in bilingual aphasia patients. *Frontiers in Human Neuroscience, 12*, 81. https://doi.org/10.3389/fnhum.2018.00081

Khateb, A., Pegna, A. J., Landis, T., Mouthon, M. S., & Annoni, J. M. (2010). On the origin of the N400 effects: An ERP waveform and source localization analysis in three matching tasks. *Brain Topography, 23*(3), 311–320. https://doi.org/10.1007/s10548-010-0149-7

Kiang, M., Patriciu, I., Roy, C., Christensen, B. K., & Zipursky, R. B. (2013). Test–retest reliability and stability of N400 effects in a word-pair semantic priming paradigm. *Clinical Neurophysiology, 124*(4), 667–674. https://doi.org/10.1016/j.clinph.2012.09.029

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Iragui, V. (1998). The N400 in a semantic categorization task across 6 decades. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, 108*(5), 456–471. https://doi.org/10.1016/S0168-5597(98)00023-9

Lew, H. L., Gray, M., & Poole, J. H. (2007). Temporal stability of auditory event-related potentials in healthy individuals and patients with traumatic brain injury. *Journal of Clinical Neurophysiology*, *24*(5), 392-397. https://doi.org/ 10.1097/WNP.0b013e31814a56e3.

MacDonald, S. W., Nyberg, L., & Bäckman, L. (2006). Intra-individual variability in behavior: Links to brain structure, neurotransmission and neuronal activity. *Trends in Neurosciences, 29*(8), 474–480. https://doi.org/10.1016/j.tins.2006.06.011

Makeig, S., Bell, A., Jung, T.-P., & Sejnowski, T. J. (1995). Independent component analysis of electroencephalographic data. *Advances in Neural Information Processing Systems*, 8.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech Language and Hearing Research, 50*, 940–967. https://doi.org/10.1044/1092-4388(2007/067)

Meechan, R. J. H., McCann, C. M., & Purdy, S. C. (2021). The electrophysiology of aphasia: A scoping review. *Clinical Neurophysiology, 132*, 3025–3034. https://doi.org/10.1016/j.clinph.2021.08.023

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal, 24*(3), 69–71.

Nicholas, L. E., & Brookshire, R. H. (1995). Presence, completeness, and accuracy of main concepts in the connected speech of non-brain-damaged adults and adults with aphasia. *Journal of Speech Language and Hearing Research, 38*, 145–156. https://doi.org/10.1044/jshr.3801.145

Nolan, H., Whelan, R., & Reilly, R. B. (2010). FASTER: Fully automated statistical thresholding for EEG artifact rejection. *Journal of Neuroscience Methods, 192*(1), 152–162. https://doi.org/10.1016/j.jneumeth.2010.07.015

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*, 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Perrin, F., & Garcıa-Larrea, L. (2003). Modulation of the N400 potential during auditory phonological/semantic interaction. *Cognitive Brain Research, 17*(1), 36–47. https://doi.org/10.1016/S0926-6410(03)00078-8

Praamstra, P., & Stegeman, D. F. (1993). Phonological effects on the auditory N400 event-related brain potential. *Cognitive Brain Research, 1*(2), 73–86. https://doi.org/10.1016/0926-6410(93)90013-U

Pulvermüller, F., Mohr, B., & Lutzenberger, W. (2004). Neurophysiological correlates of word and pseudo-word processing in well-recovered aphasics and patients with right-hemispheric stroke. *Psychophysiology, 41*(4), 584–591. https://doi.org/10.1111/j.1469-8986.2004.00188.x

Räling, R. (2016). *Age of acquisition and semantic typicality effects: Evidences for distinct processing origins from behavioural and ERP data in healthy and impaired semantic processing* [Doctoral dissertation, Universität Potsdam].

Randolph, C. (1998). *Repeatable battery for the assessment of neuropsychological status*. Psychological Corporation.

Revelle, W. (2022). psych: Procedures for Psychological, Psychometric, and Personality Research. *Northwestern University*. https://CRAN.R-project.org/package=psych.

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

Richardson, J. D., & Dalton, S. G. (2020). Main concepts for two picture description tasks: An addition to Richardson and Dalton 2016. *Aphasiology, 34*, 119–136. https://doi.org/10.1080/02687038.2018.1561417

Robson, H., Pilkington, E., Evans, L., DeLuca, V., & Keidel, J. L. (2017). Phonological and semantic processing during comprehension in Wernicke's aphasia: An N400 and phonological mapping negativity study. *Neuropsychologia, 100*, 144–154. https://doi.org/10.1016/j.neuropsychologia.2017.04.012

Silkes, J. P., & Anjum, J. (2021). The role and use of event-related potentials in aphasia: A scoping review. *Brain and Language, 219*, Article 104966. https://doi.org/10.1016/j.bandl.2021.104966

Song, Y., Zang, D. W., Jin, Y. Y., Wang, Z. J., Ni, H. Y., Yin, J. Z., & Ji, D. X. (2015). Background rhythm frequency and theta power of quantitative eeg analysis: Predictive biomarkers for cognitive impairment post–cerebral infarcts. *Clinical*

Electroencephalography and Neuroscience, 46, 142–146. https://doi.org/10.1177/1550059413517492

Šoškić, A., Jovanović, V., Styles, S. J., Kappenman, E. S., & Ković, V. (2022). How to do better N400 studies: Reproducibility, consistency and adherence to research standards in the existing literature. *Neuropsychology Review, 32*(3), 577–600. https://doi.org/10.1007/s11065-021-09513-4

Spironelli, C., & Angrilli, A. (2006). Language lateralization in phonological, semantic and orthographic tasks: A slow evoked potential study. *Behavioural Brain Research, 175*(2), 296–304. https://doi.org/10.1016/j.bbr.2006.08.038

Swick, D. (1998). Effects of prefrontal lesions on lexical processing and repetition priming: An ERP study. *Cognitive Brain Research, 7*(2), 143–157. https://doi.org/10.1016/S0926-6410(98)00019-6

Swick, D., & Knight, R. T. (1999). Contributions of prefrontal cortex to recognition memory: Electrophysiological and behavioral evidence. *Neuropsychology, 13*(2), 155. https://psycnet.apa.org/doi/10.1037/0894-4105.13.2.155.

Tanner, D., Morgan-Short, K., & Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology, 52*(8), 997–1009. https://doi.org/10.1111/psyp.12437

Wechsler, D., Coalson, D. L., & Raiford, S. E. (2008). *WAIS-IV: Wechsler Adult Intelligence Scale.* Pearson.

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers, 20*(1), 6–10.

Wilson, K. R., O'Rourke, H., Wozniak, L. A., Kostopoulos, E., Marchand, Y., & Newman, A. J. (2012). Changes in N400 topography following intensive speech language therapy for individuals with aphasia. *Brain and Language, 123*(2), 94–103. https://doi.org/10.1016/j.bandl.2012.06.005