CORTEX 203 (2026) 107–117

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Special Issue: Language disorders in understudied languages. Research article

The Arabic Discourse Assessment Tool (ADAT) can diagnose aphasia and fluency status following brain damage

Reem S.W. Alyahya [a,b,*] *and James D. Stefaniak* [c,d]

[a] *Division of Speech and Language Therapy, Department of Allied Health, School of Health and Medical Sciences, City St. George's, University of London, London, UK*
[b] *King Fahad Medical City, Riyadh, Saudi Arabia*
[c] *North London NHS Foundation Trust, London, UK*
[d] *Department of Neuropsychiatry, The National Hospital for Neurology and Neurosurgery, London, UK*

ARTICLE INFO

Article history:
Received 9 February 2026
Revised 24 June 2026
Accepted 26 June 2026
Action editor Adolfo Garcia
Published online 3 July 2026

Keywords:
Aphasia
Discourse
ADAT
Arabic
CWF
Assessments
Diagnosis

ABSTRACT

The Arabic Discourse Assessment Tool (ADAT) is a newly developed, standardised, and norm-referenced tool comprising three discourse tasks, eight high-quality discourse measures, and two transcription-less assessment approaches. ADAT can accurately differentiate discourse responses produced by people with aphasia (PwA) from those of neurotypical adults. In this study, we examined the diagnostic accuracy of ADAT in patients with brain damage. We addressed key methodological limitations in the existing literature by: (a) using ADAT to distinguish brain damaged PwA from those without aphasia; (b) diagnosing fluency status among PwA using ADAT; (c) testing the predictive validity of the tool in out-of-sample data; and (d) assessing whether transcription-less discourse assessment approaches can produce similar diagnostic accuracy to traditional, time-consuming discourse sampling. We examined 100 patients with brain damage using ADAT; 50 of whom were diagnosed with aphasia (28 fluent and 22 non-fluent aphasia). Using random-permutation nested cross-validation and bootstrapping, the study demonstrates that the combination of eight ADAT discourse measures can predict aphasia status in brain damaged patients with 90% accuracy, and the two transcription-less, clinically-friendly discourse assessment approaches can predict aphasia status with 88% accuracy. Furthermore, ADAT's discourse measures can diagnose fluency status in PwA with 81% accuracy, and the two transcription-less approaches can diagnose fluency status with 88% accuracy. The findings provide evidence that ADAT, including its discourse measures and transcription-less approaches, can be used to diagnose aphasia in brain damaged individuals and fluency status in PwA with very high accuracy, thus offering clinical applications for early identification and targeted interventions.

This article is part of a special issue entitled: Understudied languages published in Cortex.

* *Corresponding author.* Division of Speech and Language Therapy, Department of Allied Health City St. George's, University of London, London, UK.
E-mail address: rsalyahya@gmail.com (R.S.W. Alyahya).
https://doi.org/10.1016/j.cortex.2026.06.018

1. Introduction

Acquired language impairments following brain damage or neurological disorders (i.e., aphasia) impact the person's ability to comprehend, formulate, and produce language. These deficits can limit communication, social participation, quality of life and the well-being of people with aphasia (PwA). Aphasia affects up to 42% of stroke survivors (Lee et al., 2020). Clinical assessments of aphasia should provide a diagnosis of language impairment, identify communicative strengths and weaknesses, and guide intervention planning. The initial classification of aphasia is an important factor in determining the prognosis for PwA (Plowman et al., 2012). Thus, accurate diagnosis of the presence of aphasia and fluency status is crucial to ensure the patient is receiving the required rehabilitation services.

Language functions are evaluated through a range of clinical tests. Standardised clinical tests of language impairments examine a range of language skills within the main language domains of comprehension, spontaneous speech, naming, repetition, and general aphasia severity. However, a systematic review of 56 commercial and non-commercial aphasia assessments identified a critical problem in the existing literature in that none of these tests reported diagnostic data for identifying aphasia in stroke populations (Rohde et al., 2018). Instead, they tended to report data for differentiating aphasic stroke patients from neurotypical controls, which is not what these tests are used for in clinical practice. The review also found that the psychometric properties of these tests are often reported within the purchased test manuals, rather than in the peer-reviewed publications, and thereby, the diagnostic aptitudes of these tests have not been systematically evaluated (Rohde et al., 2018). The review highlighted the need for diagnostically validated aphasia tests. Furthermore, some published aphasia assessment tests have low ecological validity, and scores on these tools do not reflect naturalistic language, and they may not predict real-time performance (Fergadiotis & Wright, 2015; Marini et al., 2011). Moreover, clinical aphasia tests do not optimally assess spoken discourse and naturalistic language, despite the remarkable impact of aphasia on discourse production (Alyahya et al., 2021, 2022a; Fromm et al., 2017; Martínez-Ferreiro et al., 2017). Assessing spoken discourse is critical because it can detect higher-level expressive language deficits that might be overlooked by frequently used tests (e.g., picture naming), and thus, might contribute to more accurate diagnosis and patient-centred interventions. Moreover, discourse information provides insight into overall communicative performance, as it reflects the person's ability to exchange information, convey ideas, and participate in social interactions (Bryant et al., 2016; Dipper and Pritchard, 2017; Pritchard et al., 2017).

The Arabic Discourse Assessment Tool (ADAT) (Alyahya, 2024, 2025) was recently developed to address several limitations in existing clinical assessment tools. Firstly, published standardised aphasia assessment tools tend to be biased towards English and European languages, and the Western culture. Due to the lack of comprehensive aphasia assessment tools that are appropriate for Arabic speakers, Speech-Language Therapists/Pathologists (SLTs) in Arabic countries tend to rely on informal assessments or the use of non-adapted Western English assessment batteries (Khoja, 2019). Informal assessments should not be used for diagnosis or planning interventions. Furthermore, the use of Western English aphasia assessment tools to examine language impairment in Arabic speakers is problematic because these tests are not sensitive to the Arab culture and the linguistic features of the Arabic language. Secondly, the gold standard approach to assessing spoken discourse and naturalistic language is discourse sampling. Discourse samples provide rich data on language use and insights into various linguistic elements of spoken discourse (Bryant et al., 2016; Dipper and Pritchard, 2017). However, there is a lack of consensus on which discourse measures to use (Bryant et al., 2016). A comprehensive review of studies that assessed expressive language using discourse analysis in aphasia found 536 different discourse measures (Bryant et al., 2016). The authors argued that choosing a particular discourse measure for clinical use is highly challenging due to the lack of evidence on which measures might be most beneficial in aphasia clinical practice (Bryant et al., 2016). Moreover, there is limited evidence on the psychometric properties of most published discourse measures, and thereby the quality of these measures and their appropriateness for clinical use as diagnostic measures is doubtful (Pritchard et al., 2017). A discourse measure must be robust, and its psychometric properties of validity and reliability must be high for the measure to be considered by clinicians for use to detect deficits and characterise the features of spoken discourse in PwA. Third, discourse sampling involves the laborious process of collecting, transcribing, and analysing discourse samples. This is effortful and time-consuming, making the utilisation of discourse samples impractical in clinical settings. A survey of SLPs revealed that only 6% of clinicians always use discourse analysis as part of their clinical examination (Cruice et al., 2020). Lack of resources (tasks and norms) and time constraints were reported to be the main barriers (Bryant et al., 2017; Cruice et al., 2020).

ADAT is the first discourse assessment tool that is culturally and linguistically appropriate for use by Arabic speakers (Alyahya, 2024). The tool consists of three discourse tasks to assess descriptive, narrative, and procedural discourses using eight high-quality discourse measures that were shown to attain significantly high reliability and validity in both neurotypical controls and PwA (Alyahya, 2024). This includes *word-per-minute* (*WPM*) and *sample duration* to assess language productivity/fluency, *Lexical Information Unites* (*LIU*—both counts and percentages) to assess lexical-semantics, *number of different words* (*NDW*) to assess lexical diversity, *Correct Information Unit* (*CIU*—both counts and percentages) to assess discourse informativeness, *proportion of open and closed word classes* to assess grammatical structure, and *number of complete sentences* to assess syntactic complexity. ADAT has been further developed to incorporate two validated, clinically-friendly, transcription-less approaches that provide efficient yet accurate assessments of content word fluency, informativeness, and concepts of spoken discourse in aphasia, utilising Content Word Fluency (CWF) and Main Concept Analysis (MCA) (Alyahya, 2025). These two approaches do not require time-consuming and labour-heavy transcription or coding of discourse samples, and they can be applied in real-time while the patient is speaking to assess discourse performance during clinical examination (Alyahya, 2025). Specifically, CWF is an efficient and clinically applicable approach used to quantify the fluency of content words during spoken discourse without the need to collect, transcribe or analyse discourse responses (Alyahya et al., 2022b)· CWF consists of checklists of target items that have been developed from the discourse responses of neurotypical controls (Alyahya, 2025; Alyahya et al., 2022b). A recent neuroimaging study found lesion correlates in the left frontal and parietal regions, as well as the anterior dorsal white matter pathways in association with impaired CWF among PwA (Alyahya et al., 2026). MCA is also a discourse analysis approach used to measure the ability to produce essential information in a given discourse, accurately and completely, using a multilevel coding system of a predetermined list of key concepts (Nicholas and Brookshire, 1993, 1995).

ADAT provides an invaluable clinical resource for Arabic speakers with language impairments. Arabic is the fourth most spoken language worldwide, with over 330 million native Arabic speakers (Saad & Ashour, 2010). This highlights the importance of developing assessment tools that are linguistically and culturally suitable for use by Arabic speakers. ADAT is norm-referenced, and it can differentiate discourse performance between PwA and neurotypical controls across different discourse tasks (Alyahya, 2024). Thus, the tool has the potential to be used clinically to assess deficits in spoken discourse and help guide the development of personalised therapy objectives and strategies to improve verbal communication in PwA. However, to examine the diagnostic sensitivity of assessment tools, the tool should be tested with a representative sample of the population under study (Rohde et al., 2018), i.e., another control group of people who have suffered brain damage but without language impairments.

The present study aims to examine the diagnostic accuracy of ADAT and to determine whether its discourse measures can predict and differentiate discourse performance between people with and without aphasia in a large cohort of individuals who have sustained brain damage. We further examined the accuracy of ADAT in diagnosing fluency status in PwA following brain damage. Critically, this study performed predictive modelling analyses to test whether ADAT have predictive value in ‘out-of-sample’ data (i.e. internal cross-validation splits), thus overcoming a key limitation in the field (which has tended to identify statistical associations but not measure predictive value in ‘out-of-sample’ data). We additionally assessed whether transcription-less discourse approaches can produce similar diagnostic accuracy to the traditional, time-consuming discourse sampling.

2. Methods and materials

This study was approved by King Fahad Medical City's Institutional Review Board (IRB No. 20–763).

2.1. *Participants*

100 participants took part in this study. The inclusion criteria involved adult (above the age of 18 years), native Arabic speakers, who had previously sustained brain damage, and had self-reported normal and/or corrected-to-normal hearing/vision. All participants were medically and neurologically stable at the time of participation (i.e., at least one month post onset of brain damage). Participants were excluded from the study if they had a history of developmental speech or language difficulties before the onset of brain damage, dementia, neurodegenerative diseases, or major psychological disorders. Neurological and medical histories were obtained from patients’ medical records as documented by their neurologists.

Participants were categorised into an aphasia group and a non-aphasia group, and the aphasia group were further subcategorised into fluent or non-fluent aphasia, based on their performance on language tests (see below). In order to sample across the whole aphasia spectrum (from mild to severe aphasia), recruitment was not restricted based on aphasia severity or type. This resulted in diversity of performance within the aphasia cohort. Each group consisted of 50 participants (38 males and 18 females in each group). Among the aphasia group, 28 had fluent aphasia, and 22 had non-fluent aphasia. Both groups and sub-groups were matched on gender, age, education, and time post onset of brain damage ($p > .05$). Demographic information for both groups is reported in Table 1. All participants were recruited from King Fahad Medical City in Riyadh, Saudi Arabia, and provided written informed consent before voluntarily participating in this study according to the Declaration of Helsinki under the approval of the local research ethics committee.

2.2. *Language assessments*

Participants’ language abilities were assessed using the Short Aphasia Test for Gulf-Speakers (SATG) (Altaib et al., 2021), and the Arabic Object and Action Naming Battery (AOANB) (Alyahya & Druks, 2016). In order to be able to provide a discourse sample, non-verbal participants (i.e., who could not produce any spoken words on the SATG) were excluded from this study. Language information is reported in Table 1.

2.3. *Discourse elicitation, transcription and coding*

ADAT (Alyahya, 2024) was used to elicit discourse responses. The tool includes three tasks. The first task elicits descriptive discourse using a picture-supported stimulus called the ‘lounge composite picture description’, which involves a scene of multiple characters and events. The second task elicits narrative storytelling discourse using a picture-supported stimulus called the ‘kitchen storyboard’. The third task elicited procedural discourse by asking participants ‘how they prepare a cup of tea’. The three discourse tasks were administered in a randomised order. Picture-supported stimuli were displayed, and participants were instructed to describe what was going on in these pictures in as much detail as possible. They were allowed to look through the picture before describing it. In order for ADAT to be used in clinical practice, it was important to sample across the whole aphasia spectrum and include a range of severity levels, including severe cases, as this would reflect the standard clinical profiles of PwA. Therefore, there was no minimum number of elicited words required, and no time limit was imposed during

Table 1 – Demographic and background information for participants with brain damage.

Variables		Aphasia group (N = 5 0)			Non-aphasia groups (N = 50)			Mann–Whitney U test statistics	
		Median	IQR	Range	Median	IQR	Range	U value	P value
Demographics	**Age (years)**	48.0	23.0	18–86	46.5	20.5	18–66	1482.0	.11
	Education (years)	12.0	5.0	0–22	12.0	4.0	0–18	1048.5	.15
Medical history	**Months post onset**	6.9	16.0	1–105	10.2	21.5	1–121	1078	.31
	Medical diagnosis	Stroke = 42			Stroke = 42				
		TBI = 5			TBI = 5				
		Tumour resection = 2			Tumour resection = 2				
		Others = 1			Others = 1				
Language assessments	**AOANB (max = 100)**	52.0	57.0	0–95	97.0	4.0	91–100	13.0	<.0001
	SATG (max = 117)	83.5	35.3	17–111	113.5	4.8	103–117	31.5	<.0001

TBI = Traumatic Brain Injury, AOANB = Arabic Object and Action Naming Battery (Alyahya & Druks, 2016), SATG = Short Aphasia Test for Gulf Arabic Speakers (Altaib et al., 2021).

responses for the three discourse samples. No prompts were provided by the examiner throughout testing. Responses were continuously audio-recorded for offline transcription and analysis. Each discourse sample was transcribed verbatim (orthographically) by trained assistants.

2.4. *Statistical analyses*

Data analyses were performed using Python v3.10.12, with data handling using pandas v2.1.4 and numpy v1.26.4 and statistical analyses using statsmodels v0.14.2 and sklearn v1.3.2. Mann–Whitney *U* tests were used to determine whether continuous demographic variables (gender, age and years of education) or discourse measures were statistically significantly different between brain damaged individuals with aphasia versus those without aphasia. Statistical significance was defined as two-sided $p < .05$. Bonferroni correction was applied to the significance threshold if multiple unifactorial tests were performed on the same dataset (i.e., reported *p*-values have been left as uncorrected).

2.4.1. *Linguistic analysis*

Linguistic analysis was conducted on each transcript to code for eight high quality discourse measures that were shown to attain consistently strong psychometric properties of validity and reliability across different discourse tasks in controls and PwA (Alyahya, 2024). The discourse measures were WPM, sample duration %LIU, #CIU, NDW, proportion of open and closed word classes, and number of complete sentences. Moreover, discourse responses were further analysed to obtain scores on two transcription-less approaches: CWF, and MCA using the scoring guidelines provided by Alyahya (2025).

2.4.2. *Reliability analysis*

To determine the degree of transcription accuracy, the transcription agreement coefficient using point-by-point percent agreement was calculated on 25% of the transcripts that were randomly selected and transcribed by two independent transcribers. To examine inter-rater reliability for each discourse measure, 25% of the data was coded by another independent rater, and Intraclass Correlation Coefficient (ICC) using two-way mixed effects model and absolute agreement was computed. To estimate variance due to differences between raters, the mean square between raters was used.

2.4.3. *Predictive modelling*

Statistical associations do not necessarily imply predictive value in ‘out-of-sample’ data. This is a major limitation of many of the existing assessment tools in this field, which have tended to look at whether models are statistically significant but not proceed to test whether such models have predictive value in ‘out-of-sample’ data. This study, therefore, performed an additional set of predictive modelling analyses, which tested whether demographic data or discourse measures can be used to train logistic regression models to predict patient group (aphasia *vs* non-aphasia following brain damage) in out-of-sample individuals, significantly better than chance. ‘Out-of-sample’ is used to refer to internal cross-validation splits rather than an independent external validation cohort. Therefore, the reported performance reflects internal validation only.

Random-permutation nested cross-validation was used to provide an unbiased estimate of the mean Balanced Accuracy Score in ‘out-of-sample’ data (i.e. the generalisation error, or predictive accuracy) from logistic regression models trained on either (i) demographic data, (ii) the combination of eight discourse measures (WPM, duration, %LIU, #CIU, NDW, number of complete sentences, and proportion of open and closed class words); or (iii) the two transcription-less discourse assessment approaches (CWF and MCA) combined. The pipeline consisted of training the logistic regression model to predict patient group (aphasia *vs* non-aphasia following brain damage) of the training dataset before calculating the Balanced Accuracy Score of the trained model on ‘out-of-sample’ brain-damaged patients from the testing dataset. Similar analyses were conducted among the aphasia subgroups to predict fluency status (fluent or non-fluent aphasia). Random-permutation cross-validation was used to produce five random training/testing splits of the entire dataset (training split = 80% of the entire dataset, test split = 20% of the entire dataset), as five ‘outer-loop’ training/testing datasets. Within each ‘outer-loop’ training dataset (which in turn was 80% of the entire dataset), random-permutation cross-validation produced three random train/

test splits as three 'inner-loop' train/test datasets (training split = 66%, test split = 34%). Within the three 'inner-loop' training/testing datasets of each 'outer-loop' training dataset, hyperparameters of the logistic regression modelling algorithm (namely whether L1 or L2 regularisation was used, and the regularisation strength 'C') were optimised for that 'outer-loop' using 'GridSearchCV'.

In addition, we empirically estimated the 95% Confidence Intervals of the mean generalisation error (defined as the mean Balanced Accuracy Score in 'out-of-sample' data) from logistic regression models trained on either demographic data or discourse data, using bootstrapping (1000 iterations). Within each bootstrap iteration, the original dataset was bootstrapped to produce a 'training dataset pool' the same size as the original dataset; all samples not selected during bootstrapping for the 'training dataset pool' were included in the 'testing dataset pool'. From each 'training dataset pool', five 'outer-loop' training datasets were randomly sampled (training dataset size = .8 of the original dataset). From each 'testing dataset pool', five 'outer-loop' testing datasets were randomly sampled (testing dataset size = .2 of the original dataset). Random-permutation nested cross-validation occurred using these 'outer-loops' as above. This process was repeated on each of the 1000 bootstrap iterations to yield an empirical distribution of the mean Balanced Accuracy Score.

3. Results

The transcription agreement coefficient analysis revealed high agreement across the three discourse tasks: (i) 92% agreement for descriptive discourse; (ii) 92% agreement for narrative discourse; and (iii) 95% agreement for procedural discourse. ICC findings revealed significant excellent reliability at $p < .001$ for all discourse measures across the three discourse tasks (descriptive discourse: ICC = .976 to 1; narrative discourse: ICC = .970 to .999; procedural discourse: ICC = .942 to 1).

Predictive modelling using demographic data only as predictors was not able to predict patients within the two groups (aphasia *vs* non-aphasia following brain damage) of out-of-sample data, more accurately than chance (Table 2). This was shown by logistic regression models using demographic data having a mean (SD) cross-validation Balanced Accuracy Score of .49 (.07), and the 95% confidence interval of the mean cross-validation Balanced Accuracy Scores (.38–.63) overlapping with the 'chance' level of .5.

3.1. *The diagnostic accuracy of ADAT's discourse measures*

Predictive modelling using the discourse dataset was able to predict patient group (aphasia *vs* non-aphasia following brain damage) of out-of-sample data, significantly more accurately than chance (Table 2). Indeed, logistic regression models using the discourse dataset had a mean (SD) cross-validation Balanced Accuracy Score of .90 (.10) for all eight discourse measures (WPM, %LIU, #CIU, duration, NDW, number of complete sentences, and proportion open and closed class words), and the 95% confidence interval of the mean cross-validation Balanced Accuracy Scores (.81–.98) did not overlap with the 'chance' level of .5 (Table 2). This indicates that discourse measures alone could predict whether an 'out-of-sample' brain damaged individual had aphasia or not, with a balanced accuracy of 90%.

Furthermore, predictive modelling using the ADAT's discourse measures was able to predict fluency status (fluent *vs* non-fluent aphasia) of out-of-sample data significantly more accurately than chance, as the logistic regression models had a mean (SD) cross-validation Balanced Accuracy Score of .81 (.10) for all eight discourse measures, and the 95% confidence interval of the mean cross-validation Balanced Accuracy Scores (.57–.91) did not overlap with the 'chance' level of .5 (Table 2). This indicates that discourse measures alone can predict whether an 'out-of-sample' brain damaged patient had fluent or non-fluent aphasia with a balanced accuracy of 81%.

3.2. *The diagnostic accuracy of transcription-less discourse assessment approaches*

Predictive modelling using the two transcription-less discourse assessment approaches (CWF and MCA) was able to predict patient group (aphasia *vs* non-aphasia post brain damage) of out-of-sample data, significantly more accurately than chance (Table 2). Indeed, logistic regression models using the two transcription-less discourse assessment approaches (CWF, MCA) had a mean (SD) cross-validation Balanced Accuracy Score of .88 (.05) for the transcription-less discourse approaches (CWF and MCA), and the 95% confidence interval of the mean cross-validation Balanced Accuracy Scores (.75–.95) did not overlap with the 'chance' level of .5 (Table 2). This indicates that transcription-less assessment approaches (CWF and MCA) alone could predict whether an 'out-of-sample' brain damaged patient had aphasia or not, with a balanced accuracy of 88%.

Furthermore, predictive modelling using the two transcription-less discourse assessment approaches from ADAT was able to predict fluency status (fluent *vs* non-fluent aphasia) of out-of-sample data significantly more accurately than chance, as the logistic regression models had a mean (SD) cross-validation Balanced Accuracy Score of .88 (.14) for all eight discourse measures, and the 95% confidence interval of the mean cross-validation Balanced Accuracy Scores (.70–.99) did not overlap with the 'chance' level of .5 (Table 2). This indicates that the two transcription-less discourse approaches alone could predict whether an 'out-of-sample' brain damaged patient had fluent or non-fluent aphasia with a balanced accuracy of 88%.

3.3. *Differences in discourse performance between people with and without aphasia following brain damage*

The aphasia and non-aphasia groups of brain damaged patients differed significantly on their performance on multiple discourse measures (Table 3, Fig. 1). Specifically, at the group level, PwA produced significantly lower WPM, #CIU, %LIU, NDW, open class words ($p < .0001$) and number of complete sentences ($p = .002$) during spoken discourse with significantly

Table 2 – Predictive modelling to predict the patient's group (with aphasia *vs* without aphasia post brain damage) and patient's subgroup (fluent *vs* non-fluent aphasia) using demographic or discourse datasets.

Dependent variable	Independent variables	Mean (SD) CV Balanced Accuracy Score	95% CI of Mean CV Balanced Accuracy Score
Group (aphasia *vs* non-aphasia)	Demographic data	.49 (.07)	.38 to .63
	Discourse measures	.90 (.10)	.81 to .98
	Transcription-less discourse approaches	.88 (.05)	.75 to .95
Sub-group (fluent *vs* non-fluent aphasia)	Demographic data	.53 (.06)	.31 to .62
	Discourse measures	.81 (.10)	.57 to .91
	Transcription-less discourse approaches	.88 (.14)	.70 to .99

Demographic variables: gender, age, and years of education.
Discourse measures: WPM, sample duration, %LIU, #CIU, NDW, open class words, closed class words, and complete sentences.
Transcription-less discourse assessment approaches: CWF and MCA.

increased discourse duration ($p < .0001$) compared to the brain damaged patients without aphasia (Table 3). Interestingly, this difference was not found for one measure of grammatical complexity, as the discourse responses of the aphasia group did not have a significantly different proportion of closed-class words ($p = .67$) compared to the group of non-aphasia brain damaged patients (Table 3). Furthermore, the groups of aphasia and non-aphasia following brain damage differed significantly on their performance on both transcription-less discourse assessment approaches: PwA produced significantly lower content words (CWF) and main concepts (MCA) during spoken discourse than the non-aphasia brain damaged patients ($p < .0001$) (Table 3).

3.4. *Differences in discourse performance between fluent and non-fluent aphasia*

The discourse performance by the fluent and non-fluent aphasia groups differed significantly on several discourse measures. Specifically, the non-fluent aphasia group produced significantly fewer WPM, NDW, and open class words ($p < .0001$) than the fluent aphasia group. Furthermore, the non-fluent aphasia group produced fewer complete sentences ($p = .0004$), #CIU ($p = .003$) and %LIU ($p < .02$) than the fluent aphasia group. However, there were no differences between the two groups in the duration of discourse responses ($p = .06$), and the proportion of closed-class words ($p = .3$). Furthermore, performance on the transcription-less discourse assessment approaches differed significantly between the two groups: the fluent aphasia group produced significantly more content words (CWF) and main concepts (MCA) during spoken discourse compared to the non-fluent aphasia group ($p < .0001$).

4. Discussion

This is the first study to examine the diagnostic accuracy of a standardised aphasia assessment test within a cohort of brain damaged patients using predictive modelling. We determined the accuracy of ADAT in predicting and diagnosing brain damaged patients with and without aphasia, and in classifying aphasia patients into fluent versus non-fluent aphasia. ADAT is a recently developed novel, standardised, norm-referenced discourse assessment tool, which has been developed as a culturally and linguistically appropriate tool for use by Arabic speakers (Alyahya, 2024). The tool has been validated among neurotypical healthy controls and PwA, and eight discourse measures were shown to have strong psychometric properties and high accuracy in differentiating discourse responses produced by PwA from those produced by neurotypical healthy controls. These discourse measures assess language productivity, fluency, informativeness, lexical diversity, lexical-semantics, grammatical structure, and syntactic complexity. The present study validated ADAT further by examining its diagnostic accuracy in differentiating discourse responses produced by patients with versus without aphasia following brain damage using robust predictive modelling techniques. Three main findings were established in this study. First, predictive modelling indicated that the eight discourse measures combined can predict the patient group (whether the patient has aphasia or not following brain damage) with 90% accuracy, and can further diagnose fluency status among PwA with 81% accuracy. Second, predictive modelling also indicated that the two transcription-less discourse assessment approaches from ADAT (CWF and MCA) can predict the patient group with 88% accuracy, and can diagnose fluency status among PwA with 88% accuracy. Third, PwA produced significantly lower WPM, #CIU, %LIU, NDW, #complete sentences, proportion of open class words, content words, main concepts, and required more time to complete their discourse compared to people who did not develop aphasia after brain damage.

4.1. *Diagnostic accuracy of ADAT*

The current study provided evidence on the diagnostic accuracy of ADAT (Alyahya, 2024), and found the tool to attain very high accuracy in differentiating brain damaged patients who developed aphasia from those who did not develop aphasia using the eight discourse measures that have strong psychometric properties. Moreover, predictive modelling indicated that the two transcription-less discourse assessment approaches (CWF and MCA) of ADAT can also accurately

Table 3 – Spoken discourse performance by brain damaged patients with and without aphasia.

Discourse Measures	Proxy	Aphasia group (N = 5 0)			Non-aphasia groups (N = 50)			Mann–Whitney U test statistics	
		Median	IQR	Range	Median	IQR	Range	U value	P value
WPM	Fluency/speech rate	46.21	27.81	2.78–73.7	77.97	31.16	42.44–120.41	204	**<.0001**
Duration (seconds)	Language productivity	54.0	63.83	22.67–260.67	34.33	16.25	18.33–89.33	1929.0	**<.0001**
CIU	Informativeness	66.92	41.12	6.44–100	94.25	4.43	72.18–100	149.5	**<.0001**
LIU	Lexical-semantics	81.32	24.27	15.56–100	95.25	2.98	86.52–100	191.5	**<.0001**
NDW	Lexical diversity	17.17	9.83	2.33–67.33	31.0	16.58	14.0–67.67	490.5	**<.0001**
Complete sentences	Syntactic complexity	6.17	4.83	1.33–28.33	8.0	3.58	3.0–23.33	800.5	**.002**
Open class words (proportion)	Grammatical Category	13.33	8.67	3.67–54.0	22.0	9.17	11.0–51.33	541.0	**<.0001**
Closed class words (proportion)	Grammatical category	16.95	10.97	2.44–86.2	16.48	11.28	5.47–48.21	1311.5	.67
CWF	Shortcut to informativeness/ content word fluency	22.0	16.5	0–78	37.5	15.75	20–81	388.5	**<.0001**
MCA	Discourse concepts	14.5	10.0	0–31	27.0	4.75	19–33	138.5	**<.0001**

WPM = Words Per Minute, CIU = Correct Information Units, LIU = Lexical Information Units, NDW = Number of Different Words, CWF = Content Word Fluency, MCA = Main Concept Analysis.
Statistically significant p values are highlighted in bold (Bonferroni-corrected significance threshold of $p < .006$ due to 8 multiple comparisons).

differentiate brain damaged patients with aphasia from those without aphasia with very high accuracy. Furthermore, ADAT's discourse measures and transcription-less assessment approaches can diagnose fluency status among PwA with very high accuracy. The advantages of the two approaches over other discourse measures are their clinical applicability, as they can be administered in real-time during clinical examination without the need to record, transcribe or analyse speech samples (Alyahya, 2025). These approaches were developed to overcome challenges associated with the lengthy administration and scoring of discourse responses (Alyahya et al., 2022b; Nicholas & Brookshire, 1993). The CWF approach assesses the fluent production of content words during spoken discourse using pre-specified checklists of target words (Alyahya et al., 2022b). MCA assesses the speaker's ability to produce accurate and complete essential information of a discourse using predetermined lists of key concepts (Nicholas and Brookshire, 1993, 1995). Thus, the utilisation of the two approaches provides efficient and accurate assessment of the production of content words, concepts, and informativeness during spoken discourse.

A critical strength of this study is the methodology used to obtain these results. Specifically, rigorous predictive models were validated on out-of-sample data to ensure generalisability. The study addresses a common issue in the literature where models are often deemed successful based on statistical significance without testing their predictive power on new, unseen (i.e. out-of-sample) data. It has been shown that statistical significance in-sample does not guarantee that a model will perform well in predicting new cases (Bzdok et al., 2018). The predictive modelling method used in the current study predicted whether 'out-of-sample' individuals belonged to the aphasia or non-aphasia brain damage group based on their performance on several discourse measures. Balanced Accuracy Score was the chosen metric, as it accounts for class imbalance and provides a more meaningful evaluation of model performance in predicting group membership, and random-permutation nested cross-validation was used to ensure an unbiased estimate of the model's predictive accuracy (Varoquaux et al., 2017). Furthermore, the study utilised bootstrapping (1000 iterations) to empirically estimate the 95% confidence intervals of the mean generalisation error, providing a robust assessment of model performance variability (Efron & Tibshirani, 1994). These methods allowed us to ensure that ADAT has predictive utility in a way that is likely to be generalisable to different patients in the real world. Evaluating the performance of these models on out-of-sample data using robust methods is crucial for assessing the real-world applicability of any clinical tool.

4.2. *Discourse performance of people with versus without aphasia following brain damage*

The findings provided evidence that brain damaged patients with aphasia and those without aphasia have different performances during spoken discourse, including language productivity, fluency, informativeness, lexical-semantics, lexical diversity, and syntactic complexity. The findings indicated that PwA required more time to produce spoken discourse, and they produced significantly fewer words per minute

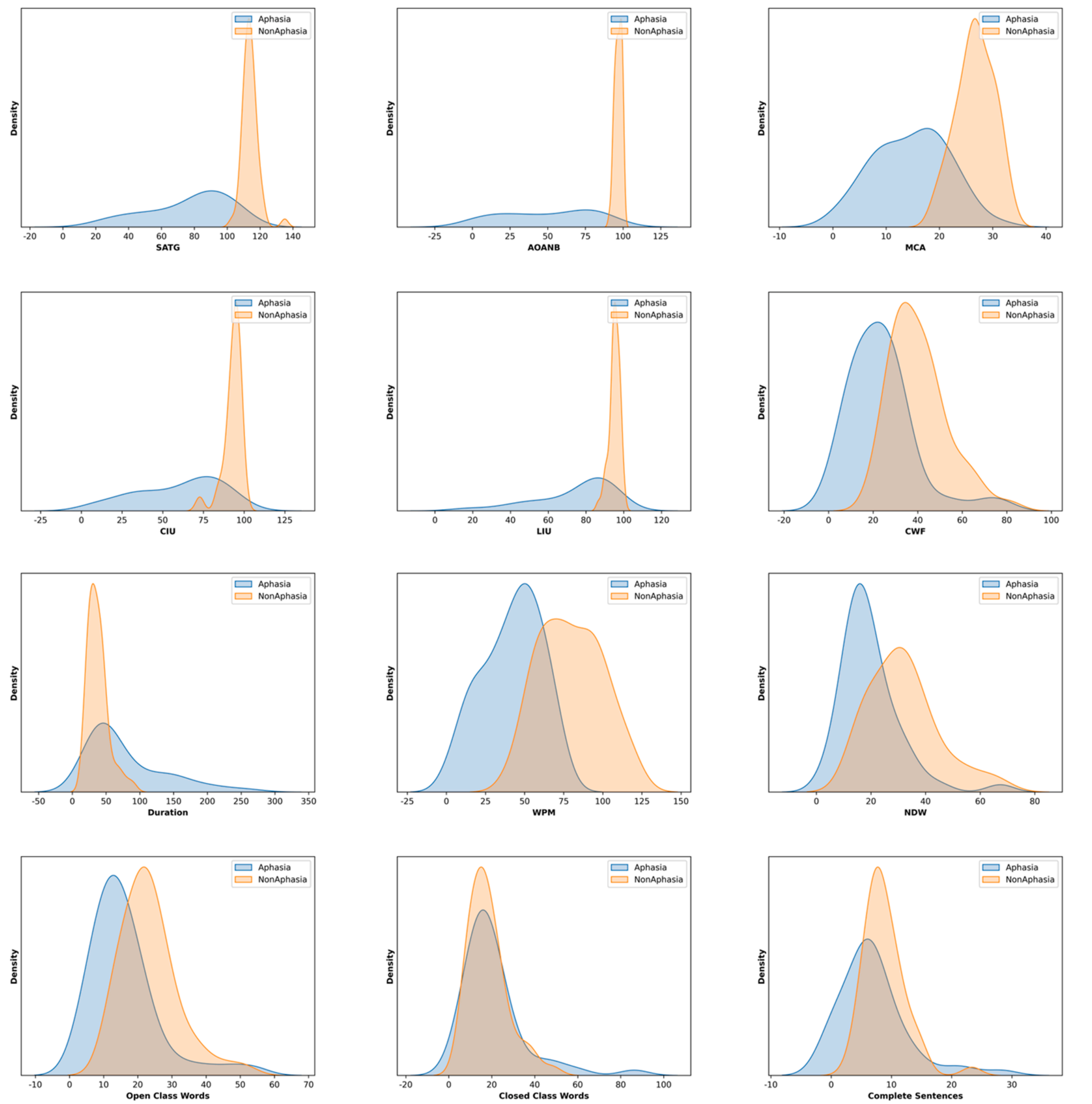

Fig. 1 – Discourse performance by brain damaged patients with and without aphasia. Density plots representing the distribution of data. The peak of the density plot shows the maximum concentration of each discourse measure in the two groups.

compared to brain damaged patients without aphasia. The increased discourse duration by the aphasia group is likely due to the presence of pauses, hesitations and self-corrections (Bello-Lepe et al., 2024). The reduced number of words produced in a minute could be due to difficulties in word retrieval (Alyahya et al., 2021). These findings suggest that fluency is affected post brain damage only in people with aphasia. The results also found that, compared to brain damaged patients without aphasia, those with aphasia produced significantly fewer units of speech that convey accurate and relevant information about the topic under discussion, and fewer meaningful and semantically accurate words. These discourse measures were even more affected among people with non-fluent aphasia compared to those with fluent aphasia. These suggest that PwA have difficulty not only with fluency but also with conveying accurate and meaningful content in their discourse, highlighting issues with semantic processing, word finding, and potentially executive function (Alyahya et al., 2022a; Kong et al., 2018). Moreover, the current study revealed that PwA used significantly fewer unique words and open class words in their discourse than brain damaged patients without aphasia, indicating reduced lexical diversity and production of words that carry the bulk of semantic content in language. Among the aphasia group, these words were produced significantly less by people with non-fluent aphasia. These highlight the challenges faced by PwA, especially those with non-fluent aphasia, in accessing a wide range of vocabulary, and particularly content words that are critical for conveying meaning during spoken discourse (Alyahya et al., 2020a; Barde et al., 2006; Fergadiotis et al., 2013). On the other hand, there were no significant differences in the use of closed class words (e.g., prepositions, conjunctions, pronouns, and articles, which serve grammatical functions in language) between brain damaged patients with and without aphasia, and also between the two aphasia groups. These interesting findings indicate that brain damaged patients with and without aphasia used closed class words at similar rates, suggesting that aphasia has a lesser effect on words that serve grammatical rather than semantic roles, possibly because these words are more automatic and have frequent usage in language (Alyahya et al., 2020b; Zhang & Hinzen, 2022). These findings also suggest that the proportion of closed class words, as a discourse measure, is less sensitive in diagnosing aphasia and fluency status post-brain damage.

4.3. *Limitations and future directions*

The study was conducted in Saudi Arabia, and thus, participants were Saudi Arabic speakers. Arabic is a diglossic language with dialectal diversity across different Arabic countries. Therefore, the Saudi dialect does not represent all Arabic dialects (e.g., Egyptian, Levantine, and North African dialects). A future direction is to adapt and validate ADAT into other Arabic dialects and other languages. ADAT can also be validated across other patient groups who might present with deficits at the discourse level, such as dementia or right hemisphere damage.

4.4. *Clinical implications*

ADAT, including its discourse measures and transcription-less discourse assessment approaches, can be used to predict the presence of aphasia following brain damage and to diagnose fluency status in PwA. This offers clinical applications for identifying at-risk patients post brain damage. Furthermore, the use of this validated tool with high diagnostic accuracy can lead to accurate diagnosis of aphasia and assessment of expressive language deficits at discourse levels, which may also lead to better clinical management.

This study provides further validation of ADAT, which benefits from strong psychometric properties, and normative data that were previously tested not only on PwA but also on an age and education matched groups of neurotypical controls. The normative data offers benchmarks of an individual's performance compared to neurotypical people, and the high predictive accuracy examined in this study provides reference points for comparison to a representative sample of people with brain damage. These comparisons allow clinicians and researchers to determine the extent to which a person with brain damage deviates from what is considered typical or expected, and whether aphasia is likely to be present or not after brain damage. This contributes to the robustness and reliability of ADAT as an assessment tool. Establishing the diagnostic accuracy of ADAT has a substantial impact on the quality of clinical management and research for PwA post brain damage or neurological disorders in Arabic countries, as well as Arabic speakers and bilingual speakers (e.g., English-Arabic, and French-Arabic) immigrants living in non-Arabic speaking countries. ADAT can replace the current problematic practice of using non-adapted English assessment tools or informal assessments with Arabic speakers with aphasia, which can result in inaccurate diagnosis and management (Altaib et al., 2021; Khoja, 2019).

Critically, it is typically challenging to collect and transcribe speech samples and then quantify discourse measures in clinical settings, due to the time-consuming nature of the traditional discourse sampling approach. It is therefore crucial that the current study demonstrates that the transcription-less discourse approaches of ADAT (CWF and MCA) offer similar diagnostic accuracy to traditional, time-consuming discourse sampling. The two transcription-less discourse assessment approaches are time-efficient, clinic-friendly, and easy to apply. Indeed, they do not require advanced linguistic skills, and thus can be used by all healthcare professionals (e.g., psychiatrists, neurologists, and nurses) in the acute phase to detect the presence of aphasia following brain damage. The current findings will therefore encourage SLTs to incorporate discourse tasks as part of their routine clinical examination, for an efficient and accurate diagnosis of aphasia, which can guide the development of personalised therapy goals and strategies to improve communication in PWA.

5. Conclusion

For the first time, we examined the diagnostic accuracy of a standardised aphasia assessment test within a cohort of brain damaged patients using predictive modelling. We addressed key gaps in the existing literature and clinical tests by: (1) examining the diagnostic accuracy of ADAT in brain damaged populations rather than differentiating PwA from neurotypical controls; and (2) testing the predictive accuracy of ADAT on out-of-sample data using predictive modelling, and moving beyond statistical significance. By employing rigorous validation techniques such as random-permutation nested cross-validation and bootstrapping, we provided robust evidence that ADAT, including its discourse measures and transcription-less discourse approaches (CWF and MCA), can reliably predict aphasia status in new patients with brain damage, thereby overcoming a common limitation in the existing literature and clinical assessments. ADAT's discourse measures can differentiate brain damage patients with aphasia from those without aphasia with 90% accuracy, and can diagnose fluency status in PwA with >81% accuracy, providing a valuable clinical assessment tool, and offering practical clinical applications for early identification and targeted interventions for individuals at risk of aphasia following brain damage. This assessment tool holds promise for improving diagnostic accuracy and tailoring rehabilitation efforts to the specific needs of patients, ultimately enhancing their communication abilities and quality of life. This study also compared the performance of brain damaged individuals who developed aphasia versus those who did not develop aphasia on several discourse measures. PwA showed significant deficits in several discourse areas, including fluency, discourse informativeness, semantics, syntactic complexity, and lexical diversity, compared to those with brain damage but without aphasia.

Data availability statement

The data are not publicly available due to ethical restrictions.

Ethics approval statement

This study was approved by King Fahad Medical City's Institutional Review Board (IRB No. 20–763).

Author contributions

RSWA: Conceptualisation, Methodology, Data curation, Investigation, Project administration, Writing – original draft, Writing – review and editing. JDS: Methodology, Formal analysis, Software, Writing – review and editing.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of interest

No conflict of interest to declare.

Acknowledgements

We are grateful to all participants for their time and effort, which made this research work possible.

Abbreviations:

AOANB	Arabic Object and Action Naming Battery
CIU	Correct Information Units
CWF	Content Word Fluency
LIU	Lexical Information Units
WPM	Words Per Minute
MCA	Main Concept Analysis
NDW	Number of Different Words
SATG	Short Aphasia Test for Gulf Arabic Speakers
PwA	People with Aphasia
TBI	Traumatic Brain Injury

REFERENCES

Altaib, M. K., Falouda, M., & Meteyard, L. (2021). From informal to formal: The preliminary psychometric evaluation of the short aphasia test for Gulf Arabic speakers (SATG). *Aphasiology, 35*, 1048–1066.

Alyahya, R. S. W. (2025). Development and real-time clinical application of new transcription-less discourse assessment approaches for Arabic-speakers with aphasia. *International Journal of Language and Communication Disorders*. https://doi.org/10.1111/1460-6984.13083. Epub ahead of print.

Alyahya, R. S. W., Conroy, P., Halai, A. D., et al. (2022b). An efficient, accurate and clinically-applicable index of content word fluency in Aphasia. *Aphasiology, 36*, 921–939.

Alyahya, R. S. W., Ding, J., Middleton, E. L., et al. (2026). The lesion correlates of impaired content word fluency during spoken discourse in aphasia. *Brain Communications, 8*. https://doi.org/10.1093/braincomms/fcag071. Epub ahead of print.

Alyahya, R. S. W., & Druks, J. (2016). The adaptation of the object and action naming battery into saudi Arabic. *Aphasiology, 30*, 463–482.

Alyahya, R. S. W., Halai, A. D., Conroy, P., et al. (2020a). Mapping psycholinguistic features to the neuropsychological and lesion profiles in aphasia. *Cortex, 124*, 260–273.

Alyahya, R. S. W., Halai, A. D., Conroy, P., et al. (2020b). A unified model of post-stroke language deficits including discourse production and their neural correlates. *Brain, 143*, 1541–1554.

Alyahya, R. S. W., Halai, A. D., Conroy, P., et al. (2021). Content word production during discourse in aphasia: Deficits in word quantity, not lexical–semantic complexity. *Journal of Cognitive Neuroscience, 33*, 2494–2511.

Alyahya, R. S. W. (2024). The development of a novel, standardized, norm-referenced Arabic discourse assessment tool (ADAT), including an examination of psychometric properties of discourse measures in aphasia. *International Journal of Language and Communication Disorders*. https://doi.org/10.1111/1460-6984.13083. Epub ahead of print 18.

Alyahya, R. S. W., Lambon Ralph, M. A., Halai, A., et al. (2022a). The cognitive and neural underpinnings of discourse coherence in post-stroke aphasia. *Brain Communications, 4*. https://doi.org/10.1093/braincomms/fcac147. Epub ahead of print.

Barde, L. H. F., Schwartz, M. F., & Boronat, C. B. (2006). Semantic weight and verb retrieval in aphasia. *Brain and Language, 97*, 266–278.

Bello-Lepe, S., Mahmood, S., Varley, R., et al. (2024). Speech pauses in speakers with and without aphasia: A usage-based approach. *Cortex, 178*, 287–298.

Bryant, L., Ferguson, A., & Spencer, E. (2016). Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics and Phonetics, 30*, 489–518.

Bryant, L., Spencer, E., & Ferguson, A. (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology, 31*, 1105–1126.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods, 15*, 233–234.
Cruice, M., Botting, N., Marshall, J., et al. (2020). UK speech and language therapists' views and reported practices of discourse analysis in aphasia rehabilitation. *International Journal of Language and Communication Disorders, 55*, 417–442.
Dipper, L., & Pritchard, M. (2017). Discourse: Assessment and therapy. In F. Fernandes (Ed.), *Advances in speech-language pathology*. InTech.
Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429246593. Epub ahead of print 15 May 1994.
Fergadiotis, G., & Wright, H. H. (2015). Modelling confrontation naming and discourse performance in aphasia. *Aphasiology, 30*, 364–380.
Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology, 22*, S397–S408.
Fromm, D., Forbes, M., Holland, A., et al. (2017). Discourse characteristics in aphasia beyond the Western Aphasia battery cutoff. *American Journal of Speech-Language Pathology, 26*, 762–768.
Khoja, M. A. (2019). A survey of formal and informal assessment procedures used by speech-language pathologists in Saudi Arabia. *Speech, Language and Hearing, 22*, 91–99.
Kong, A. P.-H., Linnik, A., Law, S.-P., et al. (2018). Measuring discourse coherence in anomic aphasia using rhetorical structure theory. *International Journal of Speech-Language Pathology, 20*, 406–421.
Lee, S., Na, Y., Tae, W.-S., et al. (2020). Clinical and neuroimaging factors associated with aphasia severity in stroke patients: Diffusion tensor imaging study. *Scientific Reports, 10*, Article 12874.
Marini, A., Andreetta, S., del Tin, S., et al. (2011). A multi-level approach to the analysis of narrative language in aphasia. *Aphasiology, 25*, 1372–1392.
Martínez-Ferreiro, S., Vares González, E., Rosell, C. V., et al. (2017). Quantifying connected discourse in Spanish-speaking individuals with aphasia: The case of mixed aphasias. *Journal of Neurolinguistics, 44*, 38–53.
Nicholas, L., & Brookshire, R. H. (1993). A system for scoring main concepts in the discourse of non-brain-damaged and aphasic speakers. *Clinical Aphasiology, 21*, 87–99.
Nicholas, L., & 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.
Plowman, E., Hentz, B., & Ellis, C. (2012). Post-stroke aphasia prognosis: A review of patient-related and stroke-related factors. *Journal of Evaluation in Clinical Practice, 18*, 689–694.
Pritchard, M., Hilari, K., Cocks, N., et al. (2017). Reviewing the quality of discourse information measures in aphasia. *International Journal of Language and Communication Disorders, 52*, 689–732.
Rohde, A., Worrall, L., Godecke, E., et al. (2018). Diagnosis of aphasia in stroke populations: A systematic review of language tests. *PLoS One, 13*, Article e0194143.
Saad, M. K., & Ashour, W. (2010). Arabic text classification using decision trees. In *he 12th international workshop on computer science and information technologies CSIT*. Moscow, Russia: Saint-Petersburg.
Varoquaux, G., Raamana, P. R., Engemann, D. A., et al. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage, 145*, 166–179.
Zhang, H., & Hinzen, W. (2022). Grammar in ‘agrammatical’ aphasia: What's intact? *PLoS One, 17*, Article e0278676.