Received: 18 August 2023 | Accepted: 30 May 2024

DOI: 10.1111/1460-6984.13083

International Journal of Language & Communication Disorders

RESEARCH REPORT

The development of a novel, standardized, norm-referenced Arabic Discourse Assessment Tool (ADAT), including an examination of psychometric properties of discourse measures in aphasia

Reem S. W. Alyahya[1,2]

[1]Department of Language and Communication Sciences, School of Health and Psychological Sciences, City, University of London, London, UK

[2]Communication and Swallowing Disorders Department, King Fahad Medical City, Riyadh, Saudi Arabia

Correspondence
Reem Alyahya, Department of Language and Communication Sciences, School of Health and Psychological Sciences, City, University of London, London, UK.
Email: rsalyahya@gmail.com

Funding information
King Fahad Medical City Intramural Research Fund, Grant/Award Number: 021-31

Abstract

Background: People with aphasia (PWA) typically exhibit deficits in spoken discourse. Discourse analysis is the gold standard approach to assess language deficits beyond sentence level. However, the available discourse assessment tools are biased towards English and European languages and Western culture. Additionally, there is a lack of consensus on which discourse measures to use and limited evidence of the psychometric properties of published discourse measurements.

Aims: (1) To develop a standardized, norm-referenced, culturally and linguistically appropriate Arabic Discourse Assessment Tool (ADAT); and (2) to examine the psychometric properties of content and construct validity and interrater reliability of different discourse measures elicited using three discourse genres (descriptive, narrative and procedural) in neurotypical control adults and matched PWA.

Methods & Procedures: Discourse samples were collected using three novel discourse stimuli that are sensitive to the Arabic language and culture from 70 neurotypical control adults and a matched group of 50 PWA. Transcription agreement was assessed. A standard approach was used to evaluate construct validity and interrater reliability for 16 discourse measures that assess fluency, language productivity, information content, lexical–semantics, lexical diversity, grammatical category, grammatical structure and syntactic complexity. Strong measures were identified based on their psychometric properties, and normative data were established on these measures. Discourse performance of PWA was then examined using the newly developed tool (ADAT).

Outcomes & Results: Transcription agreement was extremely high for all discourse stimuli in both groups. Eight discourse measures were proven to have consistently very high construct validity and consistently very good to excel-

lent reliability across the three stimuli in both neurotypical control and aphasia groups: lexical information units, content information units, words per minute, discourse duration, number of different words, number of complete sentences and proportion of open and closed class words. Norms were established on these measures, and cut-off scores of impairments were determined. Other measures showed low construct validity and variable or poor reliability across the two groups.

Conclusions & Implications: The newly developed, standardized, and norm-referenced tool (ADAT) consist of three discourse stimuli and eight high-quality discourse measures that assess multiple aspects of spoken discourse and were able to differentiate PWA from neurotypical adults consistently. ADAT also includes normative data and cut-off impairment scores. The tool has great potential to enhance clinical practice and research with Arabic speakers. Evidence was provided that not all discourse measures are of high quality, as some are vulnerable to differences between raters, discourse stimuli and groups. Clinicians and researchers can use ADAT for accurate aphasia assessments, better management plans and to monitor therapy effectiveness. ADAT can be further validated in other clinical populations with language impairments.

KEYWORDS
aphasia, Arabic, assessment, connected speech, discourse, psychometric, reliability, validity

WHAT THIS PAPER ADDS

What is already known on the subject

- Discourse analysis is the gold standard approach to assess language deficits beyond sentence level. However, existing discourse assessment tools are biased towards English and European languages and Western culture. Additionally, there is a lack of consensus on which discourse measures to use in aphasia, and limited evidence of the psychometric properties of published discourse measurements.

What this paper adds to existing knowledge

- A novel, standardized, norm-referenced Arabic Discourse Assessment Tool (ADAT) was developed and validated in this study. ADAT was further validated among PWA. The study provides evidence that not all discourse measures are of high quality and thus should not be used with confidence. Specific measures are vulnerable to the type of stimuli, the rater and/or the tested group. On the other hand, eight discourse measures were identified to be reliable between different raters and across different stimuli for the two groups, and they were able to differentiate the discourse performance of PWA from neurotypical control adults. Normative data derived from neurotypical control adults were established on these strong measures, and the performance of PWA was classified as impaired based on these norms.

What are the potential or actual clinical implications of this work?

- The present study provides a novel, standardized, norm-referenced, validated discourse assessment tool that is culturally and linguistically appropriate for use by Arabic speakers (ADAT). ADAT holds immense potential to enhance clinical practice and research with Arabic speakers. The study also identified strong discourse measures that can be used to assess language productivity, information content, lexical–semantics, lexical diversity, grammatical category, and syntactic complexity for accurate and comprehensive assessments. This will lead to better rehabilitation management by guiding the development of tailored client-centred interventions. ADAT can be utilized in clinical and research settings in PWA and has the potential to be further validated with other clinical populations.

INTRODUCTION

Acquired language impairments following brain damage or neurological disorders (i.e., aphasia) have a considerable impact on discourse production. This can limit communication with others and ultimately impacts the quality of life and well-being of people with aphasia (PWA) (Corsten et al., 2014). It is crucial to assess discourse production in aphasia for several reasons. First, discourse assessment can contribute to a more accurate diagnosis, it can guide clinicians and researchers to tailor interventions and treatment objectives to target specific areas of language impairments, and it allows clinicians to monitor how PWA respond to interventions. Second, discourse information provides insight into overall communicative performance because it reflects the person's ability to convey ideas, exchange information and participate in social interactions (Bryant et al., 2016; Dipper & Pritchard, 2017; Pritchard et al., 2017). Third, evaluating discourse in aphasia contributes to our understanding of the underlying mechanisms of language processing, shedding light on how the brain supports the construction of spoken discourse (Alyahya, 2023; Alyahya et al., 2021; Alyahya, Lambon Ralph, et al., 2022; Ding et al., 2020).

Assessment of spoken discourse in aphasia

Discourse consists of connected language samples beyond sentence and phrase levels, which is used to convey meaningful messages (Bryant et al., 2016). Producing spoken discourse is a complex task which requires the engagement of several linguistic and cognitive processes, including conceptual preparation of the intended message, activation of semantic knowledge, word retrieval, syntactic processing, phonological encoding, motor articulation, and regulating content to convey a meaningful and relevant message (Alyahya, 2023; Dell, 1983). Discourse involves the organization of coherent large units of language, and it can be used to tell a story (narrative discourse), describe an event (descriptive discourse), provide instruction (procedural discourse), express an opinion (expositional discourse), or facilitate interaction (conversational discourse) (Alyahya et al., 2020; Dipper & Pritchard, 2017; Wright et al., 2014). Recent work has shown that personal narratives are potentially beneficial in assessing and managing spoken discourse in PWA (Dipper & Cruice, 2018).

Clinical assessments of language impairments in aphasia are conducted using standardized assessment tools, such as the Boston Diagnostic Aphasia Evaluation (BDAE) (Goodglass & Kaplan, 1983), Western Aphasia Battery (WAB) (Kertesz, 1982) and Comprehensive Aphasia Test (CAT) (Swinburn et al., 2005). These tools provide an overview of deficits within main language domains, including spontaneous speech, comprehension, repetition and naming, as well as general aphasia severity. However, they do not optimally assess different aspects of discourse production (e.g., informativeness, syntactic complexity and lexical diversity). Spoken discourse in standardized aphasia assessment tools are assessed using a single composite picture description. Previous studies have provided evidence that discourse genre and stimulus type can influence the linguistic elements of spoken discourse (Alyahya et al., 2020; Olness, 2006; Schnur & Wang, 2024) Therefore, it is essential that a variety of discourse elicitation stimuli are utilized in the assessment of spoken discourse. A survey to speech–language pathologists/therapists (SLTs) revealed that only 6% of clinicians always use discourse analysis in clinical examination (Cruice et al., 2020). Another survey indicated that SLTs who utilize discourse analysis as part of language assessment in aphasia used judgement-based methods and relied heavily on clinical

observation (Bryant et al., 2017). Time constraints and lack of resources including norms and appropriate tasks were reported to be the main barriers (Bryant et al., 2017; Cruice et al., 2020).

Discourse samples are the gold standard approach to assess spoken discourse and naturalistic language; they provide rich data on language use and insights into various linguistic elements of discourse (Bryant et al., 2016; Dipper & Pritchard, 2017). However, there is a lack of consensus on which discourse measures should be used, as highlighted in a comprehensive review of studies that assessed expressive language in aphasia using discourse analysis (Bryant et al., 2016). This review found 536 different discourse measures utilized in different studies (Bryant et al., 2016). Bryant et al. (2016) argued that choosing a particular discourse measure is highly challenging because of the wealth of measures used in the aphasiology discourse literature, without any indication of which measures might be most beneficial in aphasia clinical practice. The review indicated that researchers typically select discourse measures tailored to the goals of their research rather than utilizing a body of measures that were not specific to the research questions or a specific therapy target (Bryant et al., 2016).

Discourse measures must be robust in order to be used to characterize the features of spoken discourse in PWA in clinical and research settings. Specifically, the psychometric properties of validity and reliability for each measure must be tested. The discourse measure should be grounded in relevant theory, it should measure a specific construct, and scores on the measure should be consistent across different raters. High reliability indicates that the measure can produce consistent results across different raters. Validity refers to the extent to which a measure captures important information about the underlying construct. This includes: (1) content validity, which can be evaluated by reviewing the relevance of the measure against a theoretical model; and (2) construct validity, which refers to the ability of the measure to differentiate between groups, where scores on the measure show significant differences between two groups (e.g., PWA and neurotypical persons) (Sherratt, 2007). Some aphasiology studies have evaluated the psychometric properties of spoken discourse measures in aphasia (Brookshire & Nicholas, 1994; Kong, 2009; Nicholas & Brookshire, 1993). Nevertheless, such studies are limited in number and had small sample sizes (Stark et al., 2021). Thus, it is uncertain whether these discourse measures are of sufficient quality and appropriate for clinical use (Pritchard et al., 2017). A working group was recently established to recommend standards in the collection, analysis, and reporting of spoken discourse in aphasia. Through a systematic review, this working group identified that a lack of understanding of the psychometric properties of spoken discourse measures is a current research gap. The group argued that knowing these psychometric properties would allow the identification of the best discourse measures to be used during clinical assessments (Stark et al., 2021). The limited evaluation and reporting of psychometric properties of discourse measures, is a gap in the aphasiology discourse literature that has been addressed in the current study. Obtaining discourse measures with high validity and reliability would encourage their consistent use across different studies to assess language structure at the discourse level in PWA. This would allow comparison of findings from different studies.

Arabic aphasia assessment tools

Published standardized aphasia assessment tools tend to be biased towards English and European languages and the Western culture. The standardization of such tools is essential because it allows reliable and efficient assessment procedures. Recently, there has been increasing interest in the development of aphasia assessment tools in other languages, including Arabic. There are only a few aphasia assessment tools in Gulf Arabic countries, which to date have consisted of the following:

- The Arabic Object and Action Naming Battery (AOANB) (Alyahya & Druks, 2016).
- An attempt to adapt the CAT for the use by Gulf Arabic speaking, which is currently under development (Khwaileh et al., 2016).
- The Short Aphasia Test for Gulf Arabic speakers (SATG) (Altaib et al., 2021).

Due to the lack of Arabic comprehensive aphasia assessment tools, SLTs in Arab countries tend to rely on informal tests or the use of non-adapted Western English assessments (Khoja, 2019). A survey to SLTs in Saudi Arabia indicated that 89.2% of clinicians use the BDAE or the WAB with Arabic speakers, either in their original form or, more commonly, through translating these tests without standardization, norm-referenced or cross-cultural adaptation (Khoja, 2019). The use of Western English aphasia assessment tools to examine language impairment with Arabic-speaker is problematic, as these tests are not sensitive to the linguistic features of Arabic, nor the Arab culture. Furthermore, adapting these tests for clinical use without providing norms for psychometric testing might lead to inaccurate diagnosis (Khamis-Dakwar & Froud, 2012), and clinical management plans. Arabic is the fourth most spoken language worldwide, with over 280 million native Arabic speakers (Saad & Ashour, 2010). This highlights the need to develop assessment tools that are

linguistically and culturally suitable for use by Arabic speakers.

Current study

The purpose of this study was twofold. The first aim was to develop novel, standardized, norm-referenced, culturally and linguistically appropriate Arabic Discourse Assessment Tool (ADAT). The tool includes three discourse stimuli: composite picture description, storytelling narrative, and procedural discourse. The second aim was to examine the psychometric properties of content and construct validity and interrater reliability of different discourse measures that assess fluency, language productivity, content information, lexical diversity, syntactic complexity, and grammatical structure that were elicited using ADAT in neurotypical control adults and PWA. ADAT can be used in research and clinical practice. The availability of such resources will significantly improve the clinical management of Arabic speakers with language impairments.

METHODS

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

Participants

Two groups matched on their age and education at the group level were included in this study (i.e., a neurotypical control group and an aphasia group). The control group consisted of seventy native Arabic-speaking adults with no self-reported history of neurological conditions or brain injury, and normal and/or corrected-to-normal hearing/vision (34 males; mean age = 48.24 years old (SD = 9); mean education = 12.37 years of formal education (SD = 5)). The aphasia group consisted of fifty native Arabic-speaking adults (32 males; mean age = 50.6 years old (SD = 16.3); mean education = 11.1 years of formal education (SD = 4.95)) diagnosed with aphasia post-brain damage (42 stroke, five traumatic brain injury, two post-tumour resection and one viral encephalopathy). Their aphasia was confirmed by the author using SATG (Altaib et al., 2021), and the test indicated that aphasia severity ranged from mild to severe across the group. Aphasia participants has no history of developmental speech or language difficulties prior to the onset of brain damage. Participants with concomitant motor speech disorders or cognitive disorders were excluded from the study. Aphasia participants were medically and neurologically stable at the time of participation, as reported by their neurologist (i.e., at least 1 month post-onset, mean = 16.8 months (SD = 24.5), range = 1–104). 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. 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.

The development of ADAT

Discourse elicitation stimuli

Three culturally appropriate discourse elicitation stimuli were used in this study. Two of these were picture-supported stimuli that had been developed by the author and a team of SLTs at King Fahad Medical City as part of an unpublished in-house aphasia assessment test. The pictorial images were modified by the author to improve their quality, for use in this study. No modifications were done to the characters. In the current study, the cultural appropriateness of the stimuli, in relation to their themes, events, characters, and the quality of images, were determined in two stages. Initially, five SLTs inspected the stimuli and deemed them to be culturally appreciate. Second, a focus group of 12 Arabic-speaking adults from the general public of Saudi Arabia were consulted, and they confirmed that the stimuli are culturally appropriate and did not have suggestions for modifications.

The pictorial stimuli were black-and-white line-drawings that were hand drawn by an artist. The first stimulus, called the 'Lounge', is a black-and-white line-drawing composite picture description (Figure 1) that was developed based on the theme of the 'cookie theft' picture description (Goodglass & Kaplan, 1983), which presents a multi-scene of characters and events including two incidents. The theme of the picture and all characters are appropriate to the Arab culture (e.g., drinking Arabic coffee). This stimulus was used to elicit descriptive discourse. It has been used in a previous study (Altaib et al., 2021). The second stimulus, called the 'Kitchen', is black-and-white line-drawing storyboard that was used to elicit storytelling narrative (Figure 2). The theme of this storyboard is also appropriate to the Arab culture (e.g., it depicts a woman wearing a head scarf). The third stimulus was used to elicit procedural discourse by asking participants 'how they prepare a cup of tea'.

The three stimuli were used to collect three discourse samples from all participants, both neurotypical controls and PWA. The picture-supported stimuli were kept in place

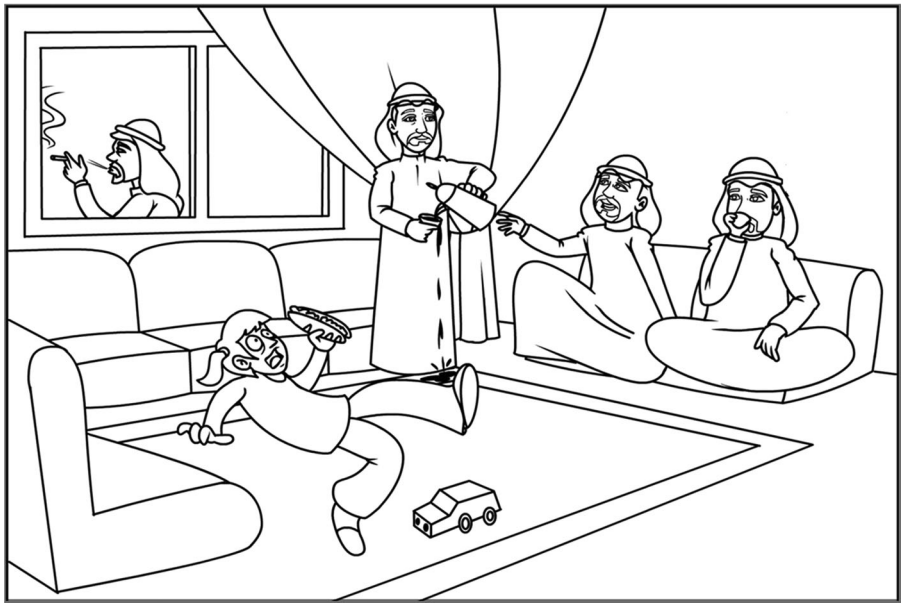

FIGURE 1 'Lounge' composite picture description stimulus.

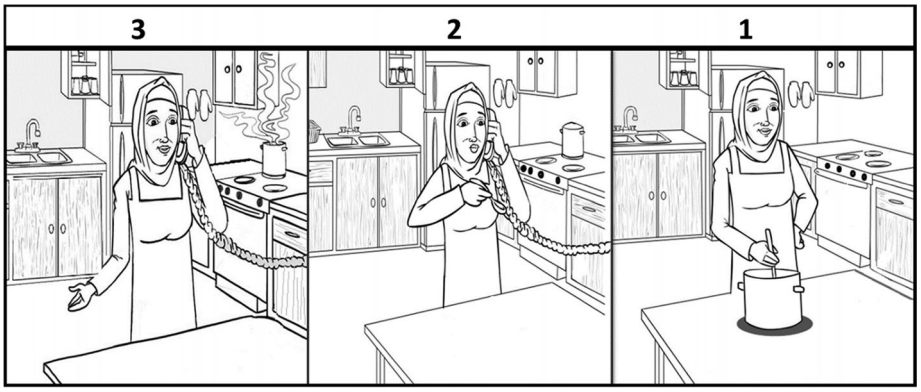

FIGURE 2 'Kitchen' storyboard narrative stimulus (right to left as in Arabic language).

and participants were asked to describe what is going on in the picture. They were allowed to look through the picture before describing it. No prompts were provided by the examiner throughout testing. For the three discourse samples, no time limit was imposed during responses. Reponses were continuously audio recorded for offline transcription and analysis.

Discourse transcriptions and analysis

Each discourse sample was transcribed verbatim (orthographically) by trained SLT assistants and checked against the recording to correct for any discrepancies. Errors were transcribed as they were heard in word form. A total of 25% of the samples were randomly selected and transcribed by another transcriber. Transcription agreement coefficient was calculated for each transcript using point-by-point percentage agreement to determine the degree of transcription accuracy for each discourse stimuli.

Linguistic analysis was conducted on each transcript to extract and code one to two discourse measures from each language domain and in relation to the structure and content of the discourse, following Bryant et al. (2016), Marini et al. (2011), Nicholas and Brookshire (1993) and Saffran et al. (1989) as follows: token counts to assess fluency, sample duration in seconds and words per minute (WPM) to assess language productivity, correct information units (CIU) as counts and percentages to assess information content, lexical information units (LIU) as counts and percentages to assess lexical–semantics, type–token ratio (TTR) and number of different words (NDW) to assess lexical diversity, proportion of nouns and verbs and proportion of open and closed word classes to assess grammatical category, number of morphemes to assess grammatical structure, mean length of utterance (MLU)

and number of complete sentences to assess syntactic complexity. These measures have been selected over other measures from the same domain (e.g., embedded sentences for syntactic complexity) because they have been frequently used in other discourse studies with PWA (Alyahya et al., 2020; Armstrong, 2001; Ding et al., 2020; Fergadiotis et al., 2013; Kong, 2009; Lau et al., 2022; Martínez-Ferreiro et al., 2017; Olness, 2006; Schnur & Wang, 2024; Wright et al., 2014).

Psychometric properties examination

To ensure content validity, all discourse measures used in this study were theory driven based on previous literature (Bryant et al., 2016; Marini et al., 2011; Nicholas & Brookshire, 1993; Saffran et al., 1989). Standard criteria for psychometric properties examination were used (Nunnally & Bernstein, 1994; Streiner & Norman, 2000). Specifically, construct validity was deemed to be high if the measure showed a significant difference between the neurotypical control and the aphasia groups, whereas very good interrater reliability was deemed if intraclass correlation coefficient (ICC) was ≥ 0.8 (Fleiss, 1971; Nunnally & Bernstein, 1994; Streiner & Norman, 2000). To examine construct validity for each measure, differences between the neurotypical control group and the aphasia group were examined using independent *t*-tests on each discourse measure. We inspected the data, and it met the assumptions of parametric statistics. A Bonferroni correction for multiple comparisons was applied to the *p*-value required for statistical significance (i.e., $p \leq 0.01$). To examine interrater reliability, 25% of the data across both groups was analysed and coded by another rater, and ICC was examined for each discourse measure using two-way mixed effects model and absolute agreement. To estimate variance due to differences between raters, mean square between raters were used. Discourse responses were collected from each participant once, as all participants attended a single session. Thus, it was not possible to examine test–retest reliability. We also examined whether the psychometric properties of validity and reliability were consistent across the three discourse stimuli (i.e., a high quality measure would not be vulnerable to discourse stimuli and groups, and would show high validity and at least very good reliability across the three discourse stimuli in both groups).

Normative data

Normative data were established for measures that showed consistently high validity and at least very good reliability across the three discourse stimuli in both groups. This involved generating descriptive statistics (mean, standard deviation, and range), and determining cut-off levels of impairment, with impairment being defined as scoring 1.5 SD below the mean of the neurotypical control group (as recommended by Brooks et al., 2011).

Validation of ADAT in aphasia

Psychometric properties examination of ADAT in PWA was conducted as described above. Furthermore, the performance of every participant from the aphasia group was classified as impaired on ADAT if they scored below the impairment cut-off. The percentage of PWA with impairments was then computed for each discourse measure across the three stimuli.

RESULTS

A total of 207 discourse samples produced by neurotypical controls were transcribed and analysed (i.e., three discourse samples from 67 participants, and two samples from the remaining three participants, as procedural discourse was not available due to technical issues). A total of 150 discourse samples by PWA were transcribed and analysed (i.e., three discourse samples from each participant). The data can be made available from the corresponding author through a reasonable request.

Transcription accuracy

To determine the degree of transcription accuracy, transcription agreement coefficient was calculated on 25% of the transcripts that were transcribed by two independent transcribers. The author independently calculated the agreement between the two sets of transcripts using point-by-point percentage agreement. The result revealed high accuracy in both groups: (1) neurotypical control group: 91% agreement for descriptive discourse; 92% agreement for narrative discourse; and 95% agreement for procedural discourse; and (2) aphasia group: 92% agreement for descriptive discourse; 92% agreement for narrative discourse; and 95% agreement for procedural discourse.

Psychometric properties

Reliability

Interrater reliability for each discourse measure across the three discourse stimuli in both neurotypical control and

aphasia groups was examined using ICC on 25% of the data that was analysed by two independent raters. ICC values and p values are reported in Table 1. The findings revealed significant, excellent reliability at $p < 0.001$ for token counts, discourse duration, LIU counts, CIU counts, %LIU, %CIU, NDW, number of morphemes, proportion of open class words, and proportion of closed class words. This excellent reliability was consistent across the three discourse stimuli (descriptive, narrative, and procedural discourses) in both control and aphasia groups. Number of complete sentences showed significant, very good reliability in the control group across the three discourse stimuli, and consistently significant excellent reliability in the aphasia group across the three discourse stimuli. The reliability for WPM was significant but not consistent, in which it was excellent for all three discourse stimuli in the aphasia group, but it varied from good to excellent reliability across the three discourse stimuli in the control group. TTR showed consistent significant very good reliability in the aphasia group, but the reliability in the control group was significant good reliability for picture description and storytelling narrative, but it was not significant for procedural discourse. The reliability for proportion of verbs was significant but varied from very good to excellent across different discourse stimuli in the two groups. On the other hand, proportion of nouns showed significant very good reliability for all discourse stimuli in the control group, and significant excellent reliability for picture description and procedural discourse only in the aphasia group. Although the reliability for MLU was significant and ranged from very good to excellent in the aphasia group, it was only good yet significant for picture description in the control groups, but it showed poor and insignificant reliability for storytelling narrative and procedural discourse in the control group.

Validity

To examine construct validity of each discourse measure, differences between the two groups on their performance on each discourse measure was examined across the three discourse stimuli. The results (Table 1 and Figure 3) revealed significant differences ($p < 0.001$) between the two groups that were consistent across all three discourse stimuli for CIU, LIU, TTR, NDW, WPM, MLU, discourse duration, and proportion of closed class words. The same results were obtained for CIU and LIU when they were measured using counts and percentages. Proportion of open class words also showed significant differences between the two groups at $p < 0.001$ for storytelling and procedural discourse, and at $p = 0.01$ for picture description. Number of complete sentences also showed significant differences between the two groups at $p < 0.001$ for picture description, $p < 0.01$ for storytelling narratives and procedural discourse. However, token counts, proportion of nouns, and number of morphemes did not show any differences between the two groups. Proportions of verbs showed significant differences between the two groups for procedural discourse only at $p < 0.001$.

In conclusion, the best measures with consistent high construct validity and consistent very good to excellent reliability across different discourse stimuli in both control and aphasia groups were LIU, CIU, WPM, discourse duration, NDW, number of complete sentences, proportion of open, and closed class words.

Normative data of ADAT

For measures that showed consistently high validity and very good to excellent reliability, normative data (Table 2) were computed for the three discourse stimuli (descriptive, narrative, procedural) derived from the neurotypical control group. Impairments cut-offs were established for these high quality discourse measures (Table 3).

Impaired discourse performance in PWA

The performance of PWA on the high quality discourse measures was deemed impaired if they scored 1.5 SD below the norms of the neurotypical control group as recommended by Brooks et al. (2011). Results are summarized in Table 3.

DISCUSSION

The present study provides and describes the development of a novel, standardized, norm-referenced, discourse assessment tool, which are culturally and linguistically appropriate for use by Arabic speakers (ADAT). The current study also examined psychometric properties of construct validity and interrater reliability for several discourse measures that assess language productivity, fluency, content information, lexical diversity, lexical–semantics, grammatical structure, and syntactic complexity across different discourse stimuli (composite picture description, picture-supported storytelling narrative, and procedural discourse) in neurotypical control adults and PWA.

The findings from this examination revealed that, for descriptive, narrative and procedural discourses in PWA, the best measures to assess (1) language productivity: WPM and sample duration rather than token counts; (2) lexical–semantics: LIU (either counts or percentages);

TABLE 1 Psychometric properties of discourse measures used to assess fluency, language productivity, content information, lexical–semantics, lexical diversity, grammatical structure and syntactic complexity in neurotypical control adults and people with aphasia.

Stimulus	Construct validity[a]			Interrater reliability[b]					
				Neurotypical control group ($N = 70$)			Aphasia group($N = 50$)		
Measure	Picture description	Storytelling narrative	Procedural discourse	Picture description	Storytelling narrative	Procedural discourse	Picture description	Storytelling narrative	Procedural discourse
Token counts	0.66	0.94	0.61	**0.997**	**0.990**	**0.997**	**0.999**	**0.999**	**0.996**
Duration	**< 0.001**	**< 0.001**	**< 0.001**	**1**	**0.952**	**0.961**	**1**	**0.997**	**1**
WPM	**< 0.001**	**< 0.001**	**< 0.001**	**0.990**	**0.868**	**0.65**	**0.997**	**0.994**	**0.998**
LIU	**< 0.001**	**< 0.001**	**< 0.001**	**0.999**	**0.996**	**0.998**	**0.995**	**0.998**	**0.942**
CIU	**< 0.001**	**< 0.001**	**< 0.001**	**0.954**	**0.917**	**0.956**	**0.991**	**0.988**	**0.994**
TTR	**< 0.001**	**< 0.001**	**< 0.001**	**0.783**	**0.743**	0.48	**0.898**	**0.909**	**0.832**
NDW	**< 0.001**	**< 0.001**	**< 0.001**	**0.962**	**0.974**	**0.967**	**0.988**	**0.987**	**0.978**
MLU	**< 0.001**	**< 0.001**	**< 0.001**	**0.561**	0.253	0.083	**0.938**	**0.866**	**0.949**
Complete sentences	**< 0.001**	**0.01**	**0.01**	**0.878**	**0.852**	**0.845**	**0.976**	**0.970**	**0.939**
Morphemes	0.048	0.5	0.3	**0.995**	**0.979**	**0.984**	**0.997**	**0.999**	**0.978**
Nouns (proportion)	0.02	0.02	0.06	**0.817**	**0.711**	**0.766**	**0.926**	0.616	**0.961**
Verbs (proportion)	0.4	0.06	**< 0.001**	**0.959**	**0.891**	**0.942**	**0.990**	**0.725**	**0.983**
Open class words (proportion)	**0.01**	**< 0.001**	**< 0.001**	**0.976**	**0.974**	**0.991**	**0.988**	**0.996**	**0.993**
Closed class words (proportion)	**< 0.001**	**< 0.001**	**< 0.001**	**0.980**	**0.949**	**0.989**	**0.993**	**0.999**	**0.988**

Note: CIU, content information units; LIU, lexical information units; MLU, mean length of utterance; NDW, number of different words; TTR, type token ratio; WPM, words per minute.

[a]Construct validity: two-tailed *p*-values of independent *t*-tests between neurotypical control and aphasia groups. Significant results at $p < 0.01$ are indicted in bold.

[b]Interrater reliability: ICC values. 1.00 = perfect reliability, 0.91–0.99 = excellent reliability, 0.81–0.90 = very good reliability; 0.61–0.80 = good reliability; 0.41–0.60 = moderate reliability; < 0.40 = poor reliability. Significant results at $p < 0.001$ are indicted in bold.

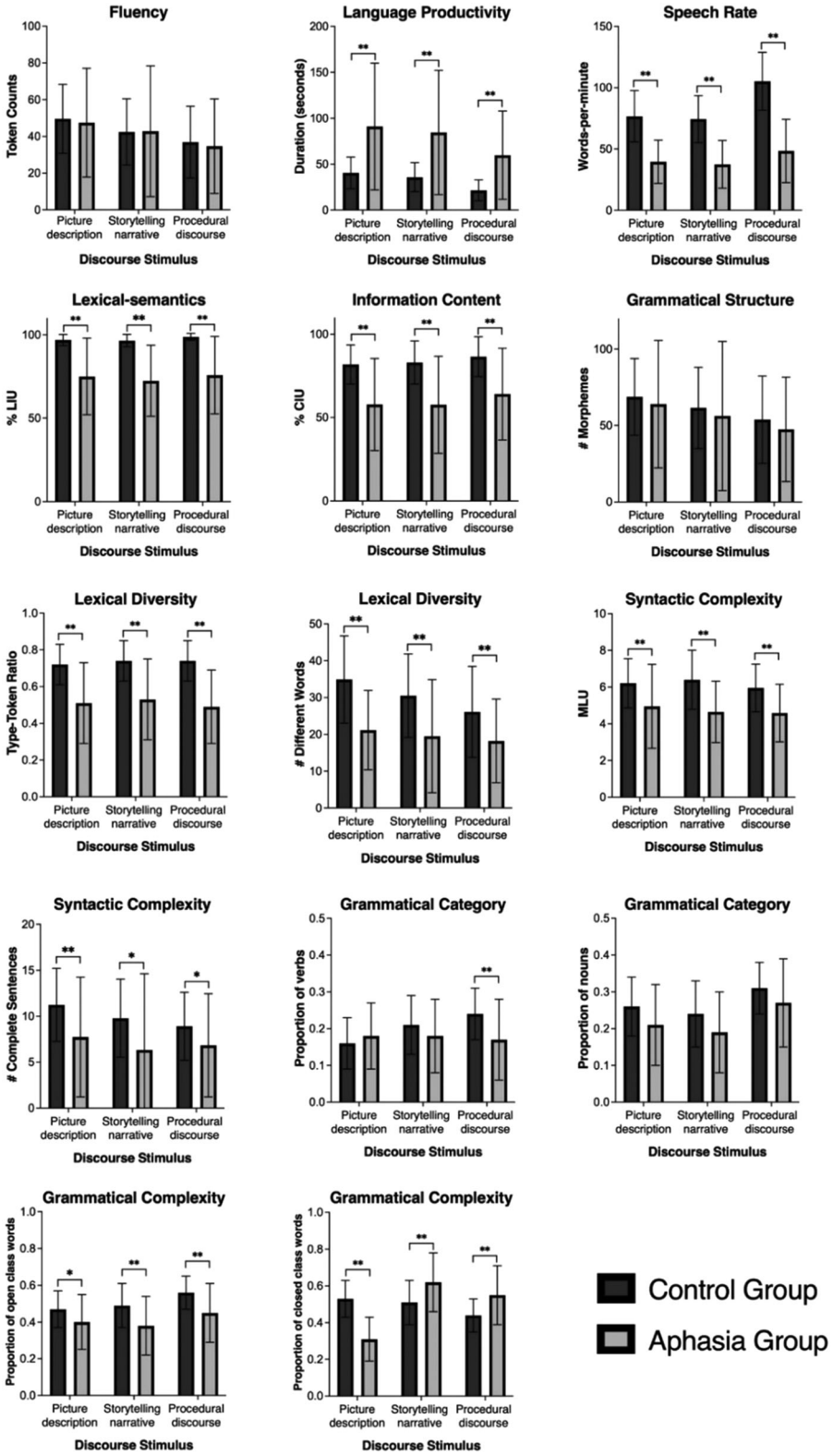

FIGURE 3 Differences between the neurotypical control group and the aphasia group on various discourse measures for three discourse stimuli.

Note: Bar graphs show the group mean and SD (errors bars). *Significant difference at $p < 0.01$, **significant difference at $p < 0.001$.

International Journal of Language & Communication Disorders

TABLE 2 Norms on high quality discourse measurements for three discourse stimuli.

Stimulus	Picture description				Storytelling narrative				Procedural discourse			
Measure	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum	Mean	SD	Minimum	Maximum
LIU (%)	96.88	3.31	87.04	100	96.56	3.69	84.13	100	98.80	2.09	92.31	100
CIU (%)	81.92	11.71	55.36	100	83.14	12.87	53.13	100	86.56	11.94	51.22	100
Duration (seconds)	40.57	17.22	12.00	101	35.85	15.80	7.00	79	21.54	11.37	5.00	56.00
WPM	76.68	20.92	28.75	141.6	74.39	19.13	35.45	115	105.22	23.65	58.38	171.4
NDW	34.90	11.93	12.00	67	30.51	11.44	8.00	57	26.12	12.44	8.00	64
Complete sentences (number)	11.21	4.00	4.00	24	9.79	4.29	4.00	25	8.91	3.73	2.00	18
Open class words (proportion)	0.47	0.10	0.24	0.67	0.49	0.12	0.24	0.80	0.56	0.09	0.33	0.92
Closed class words (proportion)	0.53	0.10	0.33	0.76	0.51	0.12	0.20	0.76	0.44	0.09	0.08	0.67

Note: CIU, content information units; LIU, lexical information units; MLU, mean length of utterance; NDW, number of different words; WPM, words per minute.

(3) information content: CIU (either counts or percentages); (4) lexical diversity: NDW rather than TTR; (5) grammatical category: proportion of open and closed word classes rather than proportion of nouns and verbs; and (6) syntactic complexity: number of complete sentences rather than MLU.

ADAT

In the current study, a novel, standardized, norm-referenced, discourse assessment tool (ADAT), was developed and validated among neurotypical Arabic-speaking adults. Furthermore, ADAT was validated in Arabic-speakers with aphasia. The tool accounts for linguistic and cultural factors, and includes culturally appropriate stimuli and norms derived from Arabic-speakers. The tool also benefits from the availability of psychometric properties of multiple discourse measures that were tested not only on PWA but also on an age and education matched neurotypical control group.

Normative data was established from the neurotypical control group on discourse measures that were deemed of high quality. Normative data is essential for any assessment tool, as it provides benchmarks and reference points for comparison of an individual's performance compared to a representative sample of neurotypical people. This comparison allows clinicians and researchers to determine the extent to which a person with language impairments deviates from what is considered typical or expected. The availability of normative data also contributes to standardized assessment procedures rather than depending on subjective judgment, ensuring that evaluations are consistent across different settings and examiners. This serves as a foundational component in developing robust and reliable assessment tools. Moreover, normative data helps establish cut-off scores (as conducted in this study) or performance ranges that would enhance diagnostic accuracy. Furthermore, normative data can guide treatment planning by providing a clear understanding of an individual's strengths and weaknesses relative to neurotypical people, which allows clinicians to develop targeted interventions that address specific areas of need.

The availability of ADAT will have a substantial impact on the quality of clinical management and research for PWA post-neurological disorders or brain damage 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 informal assessments or non-adapted English assessment tools with Arabic speakers with aphasia (Khoja, 2019), which can result in inaccurate diagnosis and management (Altaib

TABLE 3 Spoken discourse performance of people with aphasia.

Discourse		Picture description		Storytelling narrative		Procedural discourse	
Measures	**Proxy**	**Cut-off**	**Impaired**	**Cut-off**	**Impaired**	**Cut-off**	**Impaired**
WPM	Language productivity/speech rate	45	62%	46	66%	70	80%
Duration (s)	Language productivity	15	100%	13	98%	5	100%
CIU (%)	Content information	65%	56%	64%	54%	69%	50%
LIU (%)	Lexical–semantics	92%	72%	91%	82%	96%	76%
NDW	Lexical diversity	17	44%	14	40%	8	16%
Complete sentences (*n*)	Syntactic complexity	5	42%	4	46%	4	32%
Proportion of open class words	Grammatical complexity	0.32	24%	0.31	36%	0.43	44%
Proportion of closed class words	Grammatical complexity	0.38	84%	0.34	54%	0.30	24%

Note: CIU, content information units; LIU, lexical information units, NDW, number of different words; WPM, words per minute. Impaired = percentage of PWA showing impairment on this measure (i.e., scored below the cut-off).

et al., 2021; Khoja, 2019). The use of this validated tool can lead to more accurate assessment of expressive language deficits at discourse levels, which may also lead to better clinical management.

Discourse performance of PWA

Using ADAT, the performance of PWA on an individual level was dichotomized into impaired or not in comparisons to the normative data. It is interesting to notice the range of variability in the discourse performance of PWA, across the different measures reflecting the heterogenicity of this group. Specifically, the findings revealed that over two thirds of PWA had impaired language productivity and lexical–semantics; over half of PWA had impaired content information; whereas over a third of PWA had impaired lexical diversity and grammatical complexity. These results resonate with findings from other studies on PWA who speak other languages, including English (Alyahya, Conroy, et al., 2022; Alyahya et al., 2021; Armstrong, 2001; Fergadiotis et al., 2013), Chinese (Lau et al., 2022) and Spanish (Martínez-Ferreiro et al., 2017). These compelling results suggest deficits in multiple aspects of spoken discourse in aphasia irrespective of the spoken language.

The quality of discourse measurements

Discourse analysis provides rich data and insights on various linguistic elements of expressive language (Bryant et al., 2016; Dipper & Pritchard, 2017). Therefore, the quality of several discourse measures elicited during descriptive, narrative and procedural discourses in PWA and healthy controls was assessed in this study using standard psychometric criteria (Nunnally & Bernstein, 1994; Streiner & Norman, 2000). The findings revealed variability in the validity and reliability of the different discourse measures in response to the discourse stimuli and tested group. Importantly, certain measures showed variable interrater reliability that were below good on some discourse stimuli (TTR, and proportion of nouns), and across groups (TTR and MLU). Similarly, proportion of verbs showed high construct validity on one discourse stimuli but not on the other two stimuli. These inconsistencies indicate that these specific measures are vulnerable to the discourse stimulus and the tested group. Some of these measures (TTR, proportion of nouns and verbs) are derived from token counts. A further fine-grained analysis looking into verb categories and weights could provide additional insights into the linguistic features of words used by PWA during spoken discourse. Verb categories, such as state verbs (e.g., 'know', 'believe') and mental verbs (e.g., 'think', 'remember'), represent distinct semantic classes that play different roles in communication (Halliday, 2004). Verb weights, such as light verbs (e.g., 'make', 'take') and heavy verbs (e.g., 'decide', 'analyse'), are associated with different cognitive and linguistic demands (Alyahya et al., 2018). Investigating the distribution and usage patterns of these verb categories and weights among PWA could provide valuable information about their semantic processing abilities and conceptual organization.

Token count as a measure of fluency, proportion of nouns as a measure of grammatical category, and number of morphemes as measures of grammatical complexity, showed poor construct validity, as these measures were

not able to differentiate the aphasia group from the neurotypical control group. Moreover, MLU as a measure of syntactic complexity showed poor interrater reliability on all discourse stimuli in the neurotypical control group, but not with the aphasia groups. Other measures of syntactic complexity, such as embedded sentences and sentence elaboration, could yield different findings with respect to the validity and reliability of syntactic complexity measures. As these measures (token counts, TTR, MLU, number of morphemes and proportion of nouns and verbs) did not show consistently high validity and reliability across all three discourse stimuli and the two groups, they can be deemed to be low in quality, and should be avoided for clinical and research use as diagnostic or outcome measures when using picture-supported descriptive, narrative and procedural discourses with PWA.

On the other hand, other discourse measures showed extremely high validity and reliability that were consistent irrespective of the rater, discourse stimulus or group tested (LIU, CIU, WPM, discourse duration, NDW, number of complete sentences, proportion of open and closed class words). The results indicated that these discourse measures can differentiate the performance of neurotypical control adults from those with aphasia, and they can be measured reliably across different raters and different discourse stimuli. It was critical to ensure consistency of validity and reliability across different discourse measures, in order to eliminate the influence of contextual factors, as previous studies have shown that discourse stimuli can impact discourse features (Alyahya et al., 2020, 2021; Schnur & Wang, 2024), and that different discourse stimuli may tap into distinct aspects of language and cognition (Bliss & McCabe, 2006). These high quality measures can be used in clinical examination for a comprehensive assessment of language impairments, accurate diagnosis, and better planning of interventions. They also have the potential to be used as outcome measurements to monitor the effectiveness of aphasia intervention in clinical rehabilitation and in randomized control trials.

Limitations

Discourse responses in this study were audio-recorded. Future studies could also utilize videorecording during discourse assessments of PWA, which allows the observation of other means of communication including pointing or writing to augment deficits in spoken discourse, especially for people with severe aphasia. Moreover, as it is not feasible to examine all discourse measures in one study, we focused on specific discourse measures that have been commonly used in previous discourse studies (Alyahya et al., 2020; Armstrong, 2001; Ding et al., 2020; Fergadiotis et al., 2013; Kong, 2009; Lau et al., 2022; Martínez-Ferreiro et al., 2017; Olness, 2006; Wright et al., 2014). Examining the psychometric properties of other measures that have not been included in this study, and especially those related to syntactic complexity (e.g., embedded sentences), pragmatics and paralinguistics would be recommended in future studies.

Implications and future directions

The development and validation of ADAT provide a useful resource for Arabic-speakers. Moreover, the examination of the psychometric properties of several discourse measures elicited during descriptive, narrative, and procedural discourses in PWA and healthy controls identified high quality discourse measures with very high validity and reliability. This tool can be used clinically to assess deficits in spoken discourse and guide the development of personalized therapy goals and strategies to improve spoken discourse in PWA. This involves a comprehensive evaluation of the individual's communication abilities and addressing functional communication needs. Furthermore, the introduction of new validated Arabic discourse assessment tool facilitates research advancements in the field of aphasia. Researchers can investigate patterns of discourse impairment in Arabic-speaking populations, and conduct comparisons between discourse measures elicited in Arabic versus other languages. Such research contributes to the cross-linguistic understanding of aphasia, and enhances the broader knowledge base of language impairments and language processing. The utilization of these discourse measures with strong psychometric properties in neuroimaging research aiming to investigate the neural underpinning of spoken discourse, could contribute to the understanding of the underlying mechanisms of language processing, and shed light on how the brain supports the construction of spoken discourse. Furthermore, these discourse measures have the potential to be utilized in clinical and research settings to monitor therapy effectiveness by tracking progress over time and assessing changes in discourse performance in aphasia.

To examine the diagnostic sensitivity of the discourse measures to language impairments in aphasia, the psychometric properties should also be examined using a representative sample of the population under study (Rohde et al., 2018); that is, another control group of people who have suffered brain damage or neurological condition but without language impairments. This would be an avenue of future research. Another future research direction would be to examine the validity of ADAT among other clinical populations with language impairments, such as dementia.

CONCLUSIONS

The current study presents and describes the development of novel, validated, standardized, norm-referenced ADAT. This tool holds immense potential to enhance the accuracy and quality of clinical practice and research with Arabic-speakers. The tool can be used to provide a comprehensive understanding of individuals' discourse abilities, guide the planning of client-centered interventions, and monitor therapy effectiveness. Assessing spoken discourse in PWA is essential; not only because conversation between people relies heavily on producing contentful discourse, but due to the strong impact of aphasia on discourse production. The current study also established psychometric properties of construct validity and inter-rater reliability for various discourse measures elicited using ADAT. The study provided evidence that not all discourse measures are of high quality, as some are vulnerable to raters, type of stimuli, and tested groups. Thus, they cannot be used with confidence in clinical or research settings. The study also revealed the best discourse measures to be used to assess language productivity, lexical diversity, lexical–semantics, content information, grammatical category, and syntactic complexity in aphasia: LIU, CIU, WPM, discourse duration, NDW, number of complete sentences, and proportion of open and closed class words.

ACKNOWLEDGEMENTS

I am grateful to all participants for their time and effort which made this research work possible. I would like to thank Aalya AlBeeshi and Mohammad AlShalawi for their help with data collection from PWA. Many thanks to Aljoharah AlMaarik, Razan Al Towailai, and Eatedal AlHarbi for transcribing and coding discourse samples.

DATA AVAILABILITY STATEMENT

The raw data are not publicly available due to privacy and ethical restrictions. The data that support the findings of this study are available from the corresponding author upon request.

CONSENT STATEMENT

All participants 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.

ORCID

Reem S. W. Alyahya https://orcid.org/0000-0002-2766-2915

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(8), 1048–1066. https://doi.org/10.1080/02687038.2020.1765303

Alyahya, R.S.W. (2023) The structural neural correlates of spoken discourse. In: Kong, A.P.H. (Ed.) *Spoken discourse impairments in the neurogenic populations*. Springer International Publishing, Cham, pp. 111–119. https://doi.org/10.1007/978-3-031-45190-4_8

Alyahya, R.S.W., Conroy, P., Halai, A.D. & Lambon Ralph, M.A. (2022) An efficient, accurate and clinically-applicable index of content word fluency in Aphasia. *Aphasiology*, 36(8), 921–939. https://doi.org/10.1080/02687038.2021.1923946

Alyahya, R.S.W. & Druks, J. (2016) The adaptation of the object and action naming battery into Saudi Arabic. *Aphasiology*, 30(4), 463–482. https://doi.org/10.1080/02687038.2015.1070947

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

Alyahya, R.S.W., Halai, A.D., Conroy, P., & Lambon Ralph, M.A. (2018). The behavioural patterns and neural correlates of concrete and abstract verb processing in aphasia: A novel verb semantic battery. *NeuroImage: Clinical*, 17, 811–825. https://doi.org/10.1016/j.nicl.2017.12.009

Alyahya, R.S.W., Halai, A.D., Conroy, P. & Ralph, M.A.L. (2021) Content word production during discourse in aphasia: deficits in word quantity, not lexical–semantic complexity. *Journal of Cognitive Neuroscience*, 33(12), 2494–2511. https://doi.org/10.1162/jocn_a_01772

Alyahya, R.S.W., Lambon Ralph, M.A., Halai, A. & Hoffman, P. (2022) The cognitive and neural underpinnings of discourse coherence in post-stroke aphasia. *Brain Communications*, 4(3), fcac147. https://doi.org/10.1093/braincomms/fcac147

Armstrong, E. (2001) Connecting lexical patterns of verb usage with discourse meanings in aphasia. *Aphasiology*, 15(10–11), 1029–1045. https://doi.org/10.1080/02687040143000375

Bliss, L.S. & McCabe, R. (2006) Comparison of discourse genres: clinical implications. *Contemporary Issues in Communication Science and Disorders*, 33, 126–137.

Brooks, B.L., Sherman, E.M.S., Iverson, G., Iverson, G., Slick, D.J. & Strauss, E. (2011) Psychometric foundations for the interpretation of neuropsychological test results. In: Schoenberg, M.R. & Scott, J.G. (Eds.) *The black book of neuropsychology: a syndrome-based approach*. Springer Science + Business Media, pp. 893–922.

Brookshire, R.H. & Nicholas, L. (1994) Speech sample size and test–retest stability of connected speech measures for adults with aphasia. *Journal of Speech and Hearing Research*, 37(2), 399–407. https://doi.org/10.1044/jshr.3702.399

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

Bryant, L., Spencer, E. & Ferguson, A. (2017) Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology*, 31(10), 1105–1126. https://doi.org/10.1080/02687038.2016.1239013

Corsten, S., Konradi, J., Schimpf, E.J., Hardering, F. & Keilmann, A. (2014) Improving quality of life in aphasia—Evidence for the effec-

tiveness of the biographic–narrative approach. *Aphasiology*, 28(4), 440–452. https://doi.org/10.1080/02687038.2013.843154

Cruice, M., Botting, N., Marshall, J., Boyle, M., Hersh, D., Pritchard, M. & Dipper, L. (2020) UK speech and language therapists' views and reported practices of discourse analysis in aphasia rehabilitation. *International Journal of Language & Communication Disorders*, 55(3), 417–442. https://doi.org/10.1111/1460-6984.12528

Dell, G.S. (1983) A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93, 282–321.

Ding, J., Martin, R.C., Hamilton, A.C. & Schnur, T.T. (2020) Dissociation between frontal and temporal-parietal contributions to connected speech in acute stroke. *Brain*, 143(3), 862–876. https://doi.org/10.1093/brain/awaa027

Dipper, L. & Cruice, M. (2018) Personal storytelling in aphasia: a single case study of LUNA therapy. *Aphasiology*, 32(sup1), 60–61. https://doi.org/10.1080/02687038.2018.1487919

Dipper, L. & Pritchard, M. (2017) Discourse: assessment and therapy. In *Advances in speech–language pathology*. InTech.

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

Fleiss, J.L. (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378–382.

Goodglass, H. & Kaplan, E. (1983) *Boston diagnostic aphasia examination*, 2nd edition, Lea & Febiger.

Halliday, M. (2004) *An introduction to functional grammar*. Edward Arnold.

Kertesz, A. (1982) *The western aphasia battery: Test manual, stimulus cards, and test booklets*. Grune & Stratton.

Khamis-Dakwar, R. & Froud, K. (2012) Aphasia, language and culture: Arabs in the US. *Aspects of Multilingual Aphasia*, 8, 275–291. https://doi.org/10.21832/9781847697554-019

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(2), 91–99. https://doi.org/10.1080/2050571X.2017.1407620

Khwaileh, T.A., Mustafawi, E., Howard, D. & Herbert, R. (2016) An aphasia battery for Qatari/Gulf Arabic. *Front. Psychol. Conference Abstract: 54th Annual Academy of Aphasia Meeting*. https://doi.org/10.3389/conf.fpsyg.2016.68.00026

Kong, A.P.H. (2009) The use of main concept analysis to measure discourse production in Cantonese-speaking persons with aphasia: a preliminary report. *Journal of Communication Disorders*, 42(6), 442–464. https://doi.org/10.1016/j.jcomdis.2009.06.002

Lau, D.K.-Y., Kong, A.P.-H. & Chan, M.S.-W. (2022) Sentence types and complexity of spontaneous discourse productions by Cantonese-speakers with traumatic brain injury – a preliminary report. *Clinical Linguistics & Phonetics*, 36(4–5), 381–397. https://doi.org/10.1080/02699206.2021.1984582

Marini, A., Andreetta, S., del Tin, S. & Carlomagno, S. (2011) A multilevel approach to the analysis of narrative language in aphasia. *Aphasiology*, 25(11), 1372–1392. https://doi.org/10.1080/02687038.2011.584690

Martínez-Ferreiro, S., Vares González, E., Rosell Clari, V. & Bastiaanse, R. (2017) Quantifying connected discourse in Spanish-speaking individuals with aphasia: the case of mixed aphasias. *Journal of Neurolinguistics*, 44, 38–53. https://doi.org/10.1016/j.jneuroling.2017.03.001

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

Nunnally, J. & Bernstein, I. (1994) *Psychometric theory*. McGraw-Hill, New York.

Olness, G.S. (2006) Genre, verb, and coherence in picture-elicited discourse of adults with aphasia. *Aphasiology*, 20(2–4), 175–187. https://doi.org/10.1080/02687030500472710

Pritchard, M., Hilari, K., Cocks, N. & Dipper, L. (2017) Reviewing the quality of discourse information measures in aphasia. *International Journal of Language and Communication Disorders*, 52(6), 689–732. Wiley Blackwell. https://doi.org/10.1111/1460-6984.12318

Rohde, A., Worrall, L., Godecke, E., O'Halloran, R., Farrell, A. & Massey, M. (2018) Diagnosis of aphasia in stroke populations: a systematic review of language tests. *PLoS ONE*, 13(3), e0194143.

Saad, M.K. & Ashour, W. (2010) Arabic text classification using decision trees. The 12th International Workshop on Computer Science and Information Technologies CSIT.

Saffran, E.M., Berndt, R.S., Schwartz, M.F. & Saffran, E.M. (1989) The quantitative analysis of agrammatic production: procedure and data. *Brain and Language*, 37(3), 440-479.

Schnur, T.T. & Wang, S. (2024) Differences in connected speech outcomes across elicitation methods. *Aphasiology*, 38(5), 816–837. https://doi.org/10.1080/02687038.2023.2239509

Sherratt, S. (2007) Multi-level discourse analysis: a feasible approach. *Aphasiology*, 21(3–4), 375–393.

Stark, B., Dutta, M., Murray, L.L., Fromm, D., MacWhinney, B., Ramage, A.E., Roberts, A., denOuden, D.B., Brock, K., McKinney-Bock, K., Paek, E.J., Harmon, T.G., Yoon, S.O., Themistocleous, C., Yoo, H., Aveni, K., Gutierrez, S. & Sharma, S. (2021) Standardizing assessment of spoken discourse in aphasia: a working group with deliverables. *American Journal of Speech–Language Pathology*, 11(30),(1S) 491–502.

Streiner, D. & Norman, R. (2000) *Health measurement scales: a practical guide to their development and use*. Oxford University Press, Oxford.

Swinburn, K., Porter, G. & Howard, D. (2005) *The comprehensive aphasia test*. Psychology Press, Hove, UK.

Wright, H.H., Koutsoftas, A.D., Capilouto, G.J. & Fergadiotis, G. (2014) Global coherence in younger and older adults: influence of cognitive processes and discourse type. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, 21(2), 174–196. https://doi.org/10.1080/13825585.2013.794894

How to cite this article: 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 & Communication Disorders*, 1–15. https://doi.org/10.1111/1460-6984.13083