Transcription-less analysis of five discourse tasks in Laurentian French persons with post-stroke aphasia: adaptation and reliability

Amélie Brisebois, Simona Maria Brambati, Véronique Desjardins, Éva Marois, Julie Bélanger, Karine Marcotte

PII: S0010-9452(26)00143-7

DOI: https://doi.org/10.1016/j.cortex.2026.05.007

Reference: CORTEX 4363

To appear in: *Cortex*

Received Date: 20 January 2026

Revised Date: 1 May 2026

Accepted Date: 7 May 2026

Please cite this article as: Brisebois A, Brambati SM, Desjardins V, Marois É, Bélanger J, Marcotte K, Transcription-less analysis of five discourse tasks in Laurentian French persons with post-stroke aphasia: adaptation and reliability, *Cortex*, https://doi.org/10.1016/j.cortex.2026.05.007.

Transcription-less analysis of five discourse tasks in Laurentian French persons with post-stroke aphasia: adaptation and reliability

Amélie Brisebois[1,2], Simona Maria Brambati[2,3,4], Véronique Desjardins[1,2], Éva Marois[1,2], Julie Bélanger[2], and Karine Marcotte[1,2]

[1] École d'orthophonie et d'audiologie, Faculté de médecine, Université de Montréal, Montréal, Québec, Canada.

[2] Centre de recherche du Centre intégré universitaire de santé et de services sociaux du Nord-de-l'Île-de-Montréal, Montréal, Québec, Canada.

[3] Centre de recherche de l'Institut Universitaire de Gériatrie de Montréal, Montréal, Québec, Canada.

[4] Département de psychologie, Faculté des arts et des sciences, Université de Montréal, Montréal, Québec, Canada

***Correspondence:**

Amélie Brisebois, PhD

Address: Centre de recherche du Centre intégré universitaire de santé et de services sociaux du Nord-de-l'Île-de-Montréal, 5400 Gouin Ouest, Montréal, Québec, Canada, H4J 1C5

Phone number: 514-338-2222 extension 7710 Fax number: 514-340-2115

E-mail: amelie.brisebois@umontreal.ca

Keywords: aphasia, discourse, transcription-less measures, test–retest reliability

Authors email adresses: Amélie Brisebois <amelie.brisebois@umontreal.ca>; Simona Maria Brambati <simona.maria.brambati@umontreal.ca>; Véronique Desjardins <veronique.desjardins@umontreal.ca>; Éva Marois <eva.marois@umontreal.ca>; Julie

Bélanger < julie.belanger3.cnmtl@ssss.gouv.qc.ca>; Karine Marcotte <karine.marcotte@umontreal.ca>

Conflict of interest disclosure

All authors reported no conflict of interest regarding this manuscript.

Transcription-less analysis of five discourse tasks in Laurentian French persons with post-stroke aphasia: adaptation and reliability

Conflict of interest disclosure

The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.

Abstract

Discourse analysis is essential for aphasia assessment and treatment monitoring, but transcription-based methods are often too time-intensive for clinical use. Transcription-less approaches, including Main Concept Analysis (MCA), Thematic Units (TU), and coherence/reference (CoRe), offer practical alternatives, yet their psychometric properties remain underexplored. Guided by the Linguistic Underpinnings of Narrative in Aphasia framework, this study evaluated the reliability and construct validity of these measures in Laurentian French. Twenty-three French-speaking adults with chronic post-stroke aphasia completed five monologic discourse tasks across two sessions (mean interval = 11.7 days). Transcription-less measures were scored directly from audio recordings, and thirteen linguistic variables, including correct information units and moving average type–token ratio, were extracted from transcripts. Intra- and inter-rater reliabilities were assessed using intraclass correlation coefficients; test–retest reliability was examined using intraclass correlation coefficients, Spearman correlations, and Wilcoxon signed-rank tests. Construct validity was evaluated through correlations between transcription-less and linguistic (transcript-based) measures. Intra- and inter-rater reliability were good to excellent for most transcription-less measures. Test–retest reliability varied by task and measure, with CoRe showing consistently high stability across conditions. Six linguistic variables demonstrated excellent test–retest reliability. CoRe scores were

strongly associated with MC scores across tasks, and transcription-less measures correlated with information content across all tasks. MCA procedures were also adapted for three tasks in Laurentian French. This study provides the first psychometric validation of transcription-less discourse measures in Laurentian French aphasia, supporting their use as reliable, feasible, and ecologically valid tools for clinical and research discourse assessment.

Keywords: aphasia, discourse, transcription-less measures, test–retest reliability

Highlights

- Psychometric evidence supports discourse assessment in Laurentian French people with aphasia
- Transcription-less discourse measures are reliable and clinically feasible
- Coherence and reference ratings are particularly robust measures across tasks
- Culturally adapted discourse tools enable efficient tracking of discourse change

1 Introduction

Discourse, the use of language beyond the sentence level, is central to everyday communication (Armstrong, 2000). In aphasia, discourse analysis is critical for assessment and monitoring recovery across stages (Stark et al., 2021), because it captures real-world communication better than many standardized tests (Prins & Bastiaanse, 2004) and aligns with priorities expressed by people with aphasia (PWA) (Wallace et al., 2017). However, discourse research and clinical uptake remain constrained by limited standardization in elicitation and analysis, challenges that become more salient in languages other than English (Stark et al., 2021). In response, the field has increasingly emphasized systematic, psychometrically supported discourse measures (e.g., Leaman & Edmonds, 2021; Stark et al., 2023), with growing contributions from underrepresented languages, including Arabic (Alyahya, 2025), Cantonese (Kong et al., 2025), Laurentian French (Brisebois, Brambati, Jutras, et al., 2023), and Mandarin (Deng et al., 2024). This shift echoes recent calls to move toward cross-linguistically operable discourse outcomes, while addressing theoretical, methodological, and feasibility constraints that currently limit comparability across languages and sites (Sahraoui et al., 2025).

1.1 Discourse assessment

Discourse impairment is common in PWA and is a priority for assessment and treatment (Wallace et al., 2017). Because discourse measures index different communication skills, task and variable selection must be theoretically and clinically motivated (Boyle, 2020). Sampling across multiple discourse tasks (e.g., narratives, story retells, picture descriptions) improves representativeness of everyday language use (Brookshire & Nicholas, 1994a), yet many studies, including earlier work by our team (Brisebois et al., 2020, 2022), still rely on single-task protocols (Bryant et al., 2016), which can obscure multidimensional discourse profiles (Sherratt & Bryan, 2019).

Pooling discourse across tasks and/or sessions also improves measurement stability. Brookshire and Nicholas (1993) showed that variables such as words per minute and correct information units (CIUs) were more stable when combining samples across tasks, recommending ~300–400 words to balance reliability and feasibility in PWA. Subsequent work supports improved reliability when tasks are combined (Boyle, 2014; Stark et al., 2023), although the optimal task combination remains unresolved (e.g., Brisebois, Brambati, Rochon, et al., 2023; Edmonds & Babb, 2011; Whitworth et al., 2015; Zhang et al., 2020). Importantly, discourse performance varies by task (Deng et al., 2024), with tasks differentially eliciting microstructural and pragmatic features (Alyahya et al., 2020). This variability is informative but creates barriers for routine clinical use, where time, resources, and scoring complexity limit implementation (Bryant et al., 2016; Dipper et al., 2020; Boyle, 2020). These constraints motivate practical scoring approaches that reduce burden without compromising psychometric quality, including transcription-less methods designed for scalable use across tasks and populations (Stark & Dalton, 2024). This study contributes to that effort.

1.2 Advancing Discourse Assessment: Beyond Transcription

Transcription-less approaches represent a promising development in discourse assessment by enabling real-time clinical use without the burden of post-session transcription. For example, Alyahya (2025) introduced Content Word Fluency (CWF) and Main Concept Analysis (MCA) as transcription-less measures in Arabic-speaking persons with aphasia (PWA), demonstrating excellent inter-rater reliability and strong criterion and discriminative validity. However, because this work relied on a single-session design, test–retest reliability was not evaluated, leaving questions about temporal stability. More broadly, the clinical and theoretical value of transcription-less measures depends on whether they are grounded in frameworks that capture the multilevel nature of discourse processing.

The Linguistic Underpinnings of Narratives in Aphasia (LUNA; Dipper et al., 2021a) framework was developed to address this complexity. LUNA conceptualizes discourse as a dynamic system comprising four interacting levels: pragmatic, macrostructural planning, propositional organization, and linguistic formulation. Crucially, these levels are interdependent: disruptions at one level, such as lexical retrieval difficulties, can cascade to higher-order processes, affecting propositional content and global coherence. This multilevel view is well supported empirically. Prior studies have shown that discourse coherence and informativeness reflect interactions between lexical, propositional, and macrostructural processes (Marini et al., 2011a; Wright & Capilouto, 2012; Andreetta & Marini, 2015; Linnik et al., 2022).

Situating transcription-less measures within the LUNA framework clarifies both what these tools capture and how their outcomes should be interpreted. In the present study, we operationalized LUNA using three transcription-less discourse measures aligned with distinct yet interacting processing levels. Main Concept Analysis (MCA) evaluates the accuracy and completeness of essential content units, primarily indexing propositional organization while also engaging linguistic formulation (Richardson & Dalton, 2016). Thematic Units (TU) capture the production of task-relevant themes and span macrostructural planning and propositional organization (Brisebois et al., 2020). Coherence/reference (CoRe), derived from the Quality of Discourse framework (Ulatowska et al., 2003), evaluates global connectedness and referential clarity, aligning most closely with macrostructural and pragmatic levels.

Together, MCA, TU, and CoRe provide a layered yet clinically accessible approach to discourse assessment. Recent reviews have highlighted the need for further psychometric validation of transcription-less measures, particularly regarding reliability and theoretical grounding (Stark & Dalton, 2024). By embedding these measures within the LUNA framework, the present study aims to strengthen the conceptual foundation of transcription-less discourse assessment and support its

application in clinical and research contexts, particularly for French-speaking PWA.

1.3 Test-retest reliability

Evaluating test–retest reliability is a critical step in validating discourse measures, particularly transcription-less approaches intended for longitudinal research and clinical monitoring. Test–retest reliability reflects the stability of a measure over time; without it, observed changes in discourse performance may reflect measurement error rather than true change (Pritchard et al., 2018).

Discourse production is inherently variable, and reliability estimates are influenced by aphasia severity, task demands, and sample characteristics (Stark et al., 2023). Accordingly, discourse measures should be evaluated across discourse genres and with attention to known sources of variability, including lexical access, cognitive load, and sample length. While transcript-based linguistic measures, especially lexical informativeness indices such as CIUs, have been extensively examined, transcription-less discourse measures have received comparatively limited systematic evaluation.

Pritchard et al. (2018) reported mixed test–retest reliability across discourse variables, with some measures (e.g., Predicate Argument Structure) showing good stability and others (e.g., Topic and Local Coherence) demonstrating greater variability. Stark et al. (2023) extended this work by examining microlinguistic discourse measures across five monologic tasks in persons with chronic aphasia. They identified hand-scored informativeness measures, including %CIUs and CIUs per minute, as the most stable variables (often ICC > .90), particularly for longer samples. However, reliability was task-dependent; for example, CIUs per minute showed moderate reliability for the Cinderella storytelling task but excellent reliability for the Cat Rescue task. In contrast, syntactic ratios generally yielded poor-to-moderate reliability, whereas measures such as MLU and propositional density performed better in shorter samples and in individuals with more severe aphasia.

In line with Boyle (2014), Stark et al. (2023) emphasized the importance of reporting minimal detectable change (MDC) alongside ICCs, noting that reliability cannot be assumed across tasks. They also observed higher ICCs in aphasia groups than in control samples, reflecting greater between-participant variability and underscoring that reliability indices capture both measurement consistency and sample heterogeneity.

For transcription-less measures, test–retest reliability has been examined primarily for Main Concept Analysis (MCA). Across English-speaking persons with aphasia and controls (Boyle, 2014; Brookshire & Nicholas, 1994b), Cantonese-speaking populations (Kong, 2011; Kong et al., 2016), and Laurentian French-speaking controls (Brisebois, Brambati, Jutras, et al., 2023), several MCA codes, particularly Accurate and Complete (AC), Accurate and Incomplete (AI), and Absent (AB), have reached reliability thresholds recommended for research ($r > .70$; Fitzpatrick et al., 1998), with AC and AB sometimes exceeding clinical thresholds ($r > .90$; Kong, 2011). In contrast, codes reflecting inaccurate information have shown consistently poor reliability, likely due to their low frequency (Boyle, 2014).

Thematic Units (TU) have demonstrated excellent inter-rater reliability (Brisebois et al., 2020, 2022), but their test–retest reliability in persons with aphasia remains underexplored. Although TU showed poor reliability in Laurentian French-speaking controls (Marcotte et al., 2024), evidence suggests that reliability may be higher in aphasia populations (Stark et al., 2023). Similarly, while the coherence and reference (CoRe) components showed strong inter-rater agreement in their original formulation (Ulatowska et al., 2003), neither their test–retest reliability nor their use in French-speaking populations has previously been evaluated.

Overall, existing work highlights the need for systematic evaluation of transcription-less discourse measures across tasks, populations, and processing levels. Despite their clinical promise, the temporal stability of MCA, TU, and CoRe—particularly in French-

speaking persons with aphasia and within a multilevel theoretical framework—remains insufficiently documented.

1.4 Aims

The study had two primary aims: (1) to evaluate intra-rater, inter-rater, and test–retest reliability of three transcription-less discourse measures (MC, TU, and CoRe) in persons with aphasia (PWA) across five monologic discourse tasks; and (2) to examine their construct validity within the LUNA framework by testing expected relationships among these measures and with two established linguistic indicators: lexical diversity (moving average type–token ratio; MATTR) and lexical informativeness (correct information units; CIU). Secondary aims were to adapt the Main Concept Analysis (MCA) procedure for three Laurentian French discourse tasks, to estimate reliability for 13 additional linguistic discourse variables across the same tasks, and to compute minimal detectable change at the 90% confidence level (MDC_{90}) for both transcription-less and linguistic measures.

2 Materials and Methods

This project is part of a larger study and ethical approbation details appear on the title page. All the participants provided written informed consent. We report DESCRIBE Standards for Reporting Participant Characteristics in Aphasia Research (Wallace et al., 2023) and the best-practice guidelines for spoken discourse research in aphasia (Stark et al., 2022) in Supplemental Material 1.

2.1 Participants

Initial recruitment was performed between November 2023 and February 2025 in different regions of Quebec. Twenty-three participants with chronic aphasia were included: 12 females and 11 males; aged 42–84 years (M = 58.9, SD = 14.0) and education (M = 14.0, SD = 3.4). All participants performed the assessment twice, with

the interval between sessions ranging from seven to 26 days (M = 11.3, SD = 5.2). They were all right-handed and had sustained an ischemic stroke in the left hemisphere for at least 6 months prior to recruitment. No criteria concerning the initial aphasia severity were applied. No exclusion criteria were applied regarding the presence of motor speech disorders (e.g., dysarthria or apraxia of speech). All identified aphasia as the primary and most disruptive factor affecting their discourse abilities. The exclusion criteria were a history of major psychiatric disorders, learning disabilities, severe perceptual deficits (as identified by the on-call neurologist), left-handedness, and additional neurological diagnoses. This study employed a convenience sample, as recruitment timing is critical for participation. Only individuals who completed both the testing sessions were included in the final sample. Certified speech-language pathologists conducted the assessments. Severity scoring and aphasia type were based on the results obtained from the assessment tasks, clinical judgement, and overall rating on the Boston Diagnostic Aphasia Examination (BDAE) severity scale (Goodglass et al., 2001). All participants used Laurentian French daily as their primary language of communication. Seven were monolingual, eight were bilingual, and eight spoke more than two languages. This distribution reflects the linguistic diversity commonly observed in Quebec, where multilingualism, particularly among French-dominant speakers, is prevalent.

Participants' data were collected and managed using the Research Electronic Data Capture (REDCap), an electronic data capture tool hosted at the research center (Harris et al., 2009, 2019). REDCap is a secure web-based software platform that supports data capture in research studies. The participants' characteristics are shown in Table 1.

Table 1. Participants' characteristics

Participant	Age (years)	Education (years)	Sex	Time post-onset (months)	Aphasia type	Aphasia severity (BDAE scale)	Naming (TDQ30 score)	Repetition (BECLA score)	Comprehension (MT86 score)
1	61	11	M	71	Broca	Moderate	8	15	33
2	57	11	F	71	Wernicke	Moderate	20	22	40
3	69	11	M	82	Broca	Moderate to severe	7	18	38
4	84	14	M	102	Conduction	Mild to moderate	23	20	41
5	71	11	M	61	Anomic	Mild	29	23	35
6	64	12	M	67	Transcortical sensorial	Mild to moderate	20	22	39
7	48	14	F	29	Anomic	Mild	29	25	47

Participant	Age (years)	Education (years)	Sex	Time post-onset (months)	Aphasia type	Aphasia severity (BDAE scale)	Naming (TDQ30 score)	Repetition (BECLA score)	Comprehension (MT86 score)
8	54	14	F	8	Broca	Mild	24	23	40
9	51	11	M	105	Broca	Moderate	21	23	35
10	43	18	F	174	Broca	Moderate to severe	13	19	22
11	69	11	M	40	Broca	Moderate to severe	7	22	24
12	49	18	F	510	Anomic	Minimal	29	24	45
13	41	17	F	34	Anomic	Mild	27	21	46
14	63	16	F	24	Broca	Moderate	24	21	42
15	35	15	F	60	Anomic	Mild	25	25	46
16	75	13	F	156	Broca	Moderate	16	20	46

Participant	Age (years)	Education (years)	Sex	Time post-onset (months)	Aphasia type	Aphasia severity (BDAE scale)	Naming (TDQ30 score)	Repetition (BECLA score)	Comprehension (MT86 score)
17	53	9	M	13	Transcortical mixed	Moderate to severe	0	20	31
18	79	23	F	120	Broca	Moderate	14	23	51
19	70	10	F	361	Broca	Moderate	20	18	49
20	66	16	F	131	Anomic	Mild	28	25	51
21	44	14	M	39	Broca	Moderate to severe	23	25	51
22	59	17	M	60	Conduction	Moderate	22	20	47
23	39	11	M	36	Anomic	Mild	22	23	49
Mean (SD)	58.4 (13.4)	13.8 (3.4)		102.3 (116.1)			19.6 (8.0)	21.6 (2.6)	41.2 (8.2)

Participant	Age (years)	Education (years)	Sex	Time post-onset (months)	Aphasia type	Aphasia severity (BDAE scale)	Naming (TDQ30 score)	Repetition (BECLA score)	Comprehension (MT86 score)
Median [range]	59.0 [35.0–84.0]	14.0 [9.0–23.0]		67.1 [7.8–510.1]			22.0 [0.0–29.0]	22.0 [15.0–25.0]	42.0 [22.0–51.0]

Note. BDAE = Boston Diagnostic Aphasia Examination; TDQ30 = Test de dénomination de Québec - 30; BECLA = Batterie d'évaluation cognitive du langage; MT-86 = Montreal-Toulouse Aphasia Examination.

2.2 Procedure

2.2.1 Assessments

Participants completed two assessment sessions that were recorded. Discourse tasks were administered at both time points, while language and cognitive assessments varied across sessions. All assessments were conducted in Laurentian French. Testing took place in a quiet room at their local aphasia association, homes, or research center. To reach people located in areas distant from the research center, some participants performed testing in virtual sessions using the Zoom platform. At the first assessment, all participants had completed speech-language therapy and were not engaged in active language or cognitive treatment.

2.2.1.1 Language and cognitive tasks

The naming task '*Test de dénomination de Québec-30*' (TDQ30; Macoir et al., 2021), the word and pseudo-word repetition tests of the 'Batterie d'évaluation cognitive du langage '(BECLA; Macoir et al., 2015), and the Word and sentence comprehension task of the Montreal-Toulouse Aphasia Examination (MT-86; Nespoulous et al., 1992) were also conducted. Detailed language results are presented in Table 1.

2.2.1.2 *Discourse tasks*

Five monologic discourse tasks were selected for this study. Each assessment included all five spoken discourse tasks administered in a randomized order. Oral discourse tasks were: (1) a single picture description of the Picnic scene of the Western Aphasia Battery –Revised (WAB-R; Kertesz, 2006); (2) the Cinderella storytelling (Greenslade et al., 2020); (3) the Broken Window sequential picture description; (4) the Refused Umbrella sequential picture description; and (5) the Cat Rescue picture description. Instruction for the WAB-R picture description was: *'Racontez-moi ce qui se passe sur cette image'* [Tell me what is happening in the picture]. All other tasks were administered similar to the AphasiaBank protocol (e.g., MacWhinney et al., 2011). For

the Cinderella storytelling task, participants were shown wordless images in the Cinderella book and were asked to remember the story as they went on. The book was then removed, and the participants were asked to tell the story. The instruction was: '*Racontez-moi l'histoire de Cendrillon du mieux que vous pouvez*' [Tell me the Cinderella story as well as you can]. Participants were shown pictures of each task for the Broken Window, Refused Umbrella, and Cat Rescue tasks, and could refer to them while performing the task. The instruction was: *' Décrivez-moi dans vos mots et à l'aide de phrases les histoires que je vous présente. Prenez un peu de temps pour regarder les images. Elles racontent une histoire. Jetez un coup d'œil à chacune d'elles, puis je vous demanderai de me raconter l'histoire avec un début, un milieu et une fin. Vous pouvez regarder les images pendant que vous racontez l'histoire.*' [Describe in your own words, using complete sentences, the stories I am presenting to you. Take a moment to look at the images. They tell a story. Take a look at each of them, then I will ask you to tell me the story with a beginning, a middle, and an end. You may look at the images while you are telling the story] (Main Concept Analysis Training Materials v03.18.2021; Richardson & Dalton, 2021). Twenty participants performed all discourse tasks, as three did not complete the Cinderella storytelling for various reasons (e.g., unfamiliar with the story prior to the task, were too tired). The examiner was instructed not to interrupt the participant and to encourage elaboration if needed.

2.3 MC Adaptation

The main Concept (MC) adaptation was conducted for the Broken Window, Refused Umbrella, and Cat Rescue tasks. For the Cinderella storytelling task, the MCs have already been adapted to Laurentian French (Brisebois et al., 2023). The MCs for the Broken Window sequential picture description were translated and adapted from Richardson and Dalton (2016), while those for the Refused Umbrella and Cat Rescue tasks were adapted from Richardson and Dalton (2019). These three tasks were selected for their cultural relevance to Laurentian French speakers, who share a

broadly similar cultural context with American English speakers regarding the stimuli used. The completeness and accuracy of the main concepts were scored using the system developed by Richardson and Dalton (2016). The first aspect was the presence or absence of each main concept (AB). If present, the concept receives one of the following four codes: accurate and complete (AC), accurate but incomplete (AI), inaccurate but complete (IC), and inaccurate and incomplete (II). The AC, AI, IC, and II codes allow the examiner to analyze the quality of the information and provide more details on the overall informativeness. MC (total composite score of all Main Concepts) was computed according to Richardson and Dalton's (2016) formula (MC = (3 × AC) + (2 × AI) + (2 × IC) + (1 × II)). Table 2 defines each code and its corresponding scoring values.

The adaptation process for these MC checklists involved translating them into Laurentian French and refining the scoring protocol. Like previous adaptations such as those by Criel et al. (2021) and Yazu et al. (2021), our approach aimed to ensure linguistic and cultural appropriateness. Initially, we used DeepL Translator (DeepL Traduction – DeepL Translate, 2024) to generate a French MC version. A first-year student in a professional master's program in speech-language pathology, a native speaker of Laurentian French with advanced proficiency in English, reviewed the drafts to ensure semantic consistency with the original versions. Final adjustments were made through discussions among the authors. The final MC list for these three tasks is provided in Supplemental Material 2.

Table 2. Richardson and Dalton's (2016) Main Concept scoring system

Label	Score for each MC	Definition
Accurate and Complete (AC)	3 points	The statements contain all the correct information.
Accurate and Incomplete (AI)	2 points	The statements contain correct pieces of information but fail to include one essential element.
Inaccurate and Complete (IC)	2 points	The statements contain at least one incorrect piece of information but mention all essential elements.
Inaccurate and Incomplete (II)	1 point	The statements contain at least one incorrect essential element and fail to include at least one essential element.
Absent (AB)	0 points	The statements are absent.

The MC Composite (total composite score of all MCs) was computed according to Richardson and Dalton's (2016) formula (MC = (3 × AC) + (2 × AI) + (2 × IC) + (1 × II)).

2.4 Data collection and sample analysis

Discourse samples were recorded using an iPad when the session was in person or using Zoom when the session was virtual.

2.4.1 Scoring of transcription-less variables

Audio recordings of each discourse sample were extracted to score the transcription-less measures. These variables were rated in a manner designed to approximate the real-world clinical conditions. Judges listened to an audio recording of each task twice while scoring TUs, MC, and CoRe on a Microsoft Excel spreadsheet. The audio files were used to score the MCs manually using Microsoft Excel. Raters (first and third authors and a research assistant) were trained on approximately 10 samples separately (which were not included in the analyses) and discussed potential issues before performing the final scoring of the first assessment (test). Since the second round of scoring for intra-rater agreement occurred approximately five weeks apart, no refresher sessions were needed. Two raters (the first author and a research assistant) listened twice to each audio sample for each MC list. The method was the same for the TU list of the picture description of the WAB-R and the CoRe scores for each task.

2.4.1.1 Thematic Units

Thematic Units (TUs) consist of a finite list of items specific to a stimulus and were developed for the picture description task of the WAB (Brisebois et al., 2020). More specifically, 16 TUs produced by at least 75% of 45 healthy Laurentian French speakers who completed the picture description of the WAB were identified. TUs were scored only for the picture description task in the present study. This measure refers to the number of specific units the participants produced, and a maximum of 16 TUs were obtained. Following the same rules, 16 TUs were included in the analysis grid and scored as one point. The efficiency of thematic Units was also measured by calculating

the number of TUs per minute. The TU are listed in Table 3.

Table 3. Thematic Units for the picture description of the Western Aphasia Battery - Revised (Brisebois et al., 2020).

Thematic Units	
Bateau (boat)	*Lac/rivière/mer/eau* (lake/ river/ sea/ water)
Cerf-volant (kite)	*Lire* (to read)
Château de sable/sable (sand castle/sand)	*Maison/chalet* (house. country house)
Chien (dog)	*Pêcher/pêcheur* (to fish/fisherman)
Femme/madame/maman (woman/Mrs./mom)	*Pique-nique* (picnic)
Fille (girl)	*Radio/écouter de la musique* (radio/to listen to music)
Garçon (boy)	*Boire/breuvage* (to drink/ beverage)
Homme/monsieur/papa (man/Mr./ dad)	*Voiture/auto* (car)

2.4.1.2 Main Concepts

MC scoring was performed using training materials and scoring guidelines (Richardson & Dalton, 2016) provided on the AphasiaBank website (https://aphasia.talkbank.org/discourse/MainConcepts/), including video training sessions.

2.4.1.3 Coherence and reference

Ulatowska et al. (2003) introduced the components of coherence and reference (CoRe) as part of a broader framework for evaluating discourse quality. Discourse quality reflects multiple levels of information processing within the discourse macrostructure. Among its core elements, coherence refers to the maintenance of a central theme and assesses how logical ideas are connected throughout a narrative (Kintsch & van Dijk, 1978), whereas reference pertains to the clarity and precision with which story elements are introduced and tracked. By contrast, emplotment, which involves structuring events into a cohesive narrative, is specific to narrative discourse. Given its narrative specificity, emplotment was not included in this study. Instead, this dimension is addressed by scoring Thematic Units and Main Concepts that focus on task-relevant information. The scoring guidelines for CoRe are listed in Table 4.

For all tasks (i.e., WAB-R picture description, Cinderella storytelling, Broken Window, Refused Umbrella, Cat Rescue), coherence, and reference were each rated on a 5-point scale (0–4), allowing for a maximum total CoRe score of 8 points for each task.

Table 4. Coherence/reference (CoRe) scoring system (Ulatowska et al., 2003), including adapted scoring in Laurentian French.

Coherence	Score	Definition
	4 points	*Toutes les parties de la réponse sont interconnectées et claires.* All portions of the response are interconnected and clear.
	3 points	*La majeure partie de la réponse est connectée et claire, avec quelques problèmes.* Most of the response is connected and clear, with some problems.
	2 points	*Certains éléments de la réponse sont connectés.* Some of the elements of the response are connected.
	1 point	*Le discours n'est pas interprétable.* The discourse is not interpretable.
	0 point	*Aucune réponse* No response
Reference	4 points	*Tous les référents et la relation entre eux sont clairs.* All referents and the relationship between them are clear.
	3 points	*Quelques erreurs de référence* Some reference errors
	2 points	*Beaucoup d'erreurs de référence* Many reference errors
	1 point	*Aucun des référents, ni leur relation, n'est interprétable* None of the referents, nor their relationship, is interpretable
	0 point	*Aucune réponse* No response

2.4.2 Linguistic variables: transcriptions and extraction

Linguistic variables were extracted to assess construct validity. To perform linguistic analysis, video and audio files of each discourse sample were imported and transcribed by the EUDICO Language Annotator (ELAN; Sloetjes & Wittenburg, 2008). The recordings were fully transcribed orthographically. Experienced speech-language pathologists (the first author) and a second-year student in a professional master's program in speech-language pathology (the third author) transcribed the samples using the Code for the Human Analysis of Transcripts (CHAT) conventions (MacWhinney, 2000) with additional guidelines for French users (Colin & Le Meur, 2016). Utterance segmentation was performed using the CHAT conventions (MacWhinney, 2000) and was based on a combination of phonological, syntactic, and semantic criteria (see also Marini et al., 2011b). The same transcriber transcribed both the test and retest samples from the same participant for consistency.

Once the transcription was completed, morphological and grammatical information coding was conducted using the CLAN program *mor*, which tags morphemes and words under each utterance in the transcripts. Subsequently, all microstructural variables were extracted for each sample using the EVAL program of CLAN (MacWhinney 2000). MATTR was extracted with the command 'freq +tPAR +b10 +d3, the CIU count was automatically extracted using the command ' freq +t*PAR +d2 , which extracts only the CIUs, and ' freq +t*PAR +r6 +d2' was used to extract the total number of words.

2.5 Dependent variables

2.5.1 Transcription-less variables

For the picture description of the WAB-R, TU and the derived efficiency measure (TU/min) were used, following the procedure outlined in Brisebois et al. (2020). For the four other tasks, we used the MC scoring system adapted to Laurentian French

(Brisebois et al., 2023). This study used Richardson and Dalton's MC scoring system (2016), as shown in Table 2. The variables were MC, AC, AI, IC, II, and AB. We also used the derived efficiency measure (MC per minute, MC/minute). The CoRe was used for all five tasks.

2.5.2 Linguistic variables

The initial selection of the linguistic variables was based on Stark (2019). These variables are described in Table 5 and include the mean length of utterance (MLU), sample duration, propositional density (Fromm et al., 2016), number of words per minute, number of verbs per utterance, open-closed class ratio, noun-to-verb ratio, number of tokens, Correct Information Units (CIUs; Nicholas & Brookshire, 1993), percentage of CIUs, and moving average token-type ratio (MATTR; Covington, 2007).

Table 5. Definition of the linguistic variables.

Measure	Definition	Language dimension
Duration	Duration of the sample in seconds	Corpus size
Tokens	Total number of words produced	Corpus size
Mean length of utterance (MLU)	Average number of words per utterance	Productivity
Propositional density	Number of verbs, adjectives, adverbs, prepositions and conjunctions divided by the total number of words	Content richness
Words per minute	Total number of tokens divided by the duration (converted from seconds to minute)	Fluency
Verbs per utterance	Average number of verbs (verbs, copulas, auxiliaries followed by past or present participles) per utterance.	Syntactic complexity
Open/closed class ratio	Ratio of open class words (all nouns, verbs, copulas, adjectives and adverbs) divided by closed class words (all other words)	Syntactic complexity
Noun/verb ratio	Ratio of nouns to verbs, excluding auxiliaries and modals	Syntactic complexity

Measure	Definition	Language dimension
Moving Average Token-Type Ratio (MATTR)	Average of estimated Token-Type Ratios for successive nonoverlapping successive windows of flixed length	Lexical diversity
% Correct information units (CIUs)	Total number of words relevant to the stimulus and informative (CIUs) divided by the total number of words	Lexical informativeness
CIU per minute	Total number of CIUs divided by the duration (converted from seconds to minute)	Lexical informativeness

Note. Data derived from the CLAN software (MacWhinney et al., 2010).

2.6 Data analysis

2.6.1 Intra and inter-rater reliability of the transcription-less measures

To determine intra- and inter-rater reliability in TU, MC, and CoRe scoring, samples from six participants per rater (representing approximately 26% of the samples; a total of 30 samples) were randomly selected for each rater. For intra-rater reliability, raters scored each measure twice, with 14–77 days between scores (mean = 26.4; SD = 14.8 days). For inter-rater reliability, the third author and a research assistant initially scored the samples and the first author scored a second time the same samples.

2.6.2 Inter-rater reliability of the transcripts

To determine inter-rater reliability in transcription, 25 transcripts per rater (representing 21% of the transcripts each) were randomly selected for each rater. Precisely, the third author transcribed samples were initially transcribed by the first author and vice versa. The total number of tokens represents the transcription accuracy. The number of utterances is critical in the CHAT format because it relies uniquely on the transcriber's competence to distinguish the utterance boundaries. The reliability of this measure suggests consistency in utterance segmentation throughout the samples.

2.7 Statistical analyses

2.7.1 Task combination for analyses

This manuscript presents separate intra-rater, inter-rater, and test-retest reliability analyses for the WAB-R and Cinderella storytelling tasks and an analysis for the combined tasks of the Broken Window, Refused Umbrella, and Cat Rescue. Both methodological and psychometric considerations have informed this decision. The three combined tasks tend to elicit shorter discourse samples, which limits their suitability for standalone test–retest reliability analysis. Given the dual aim of this

manuscript, to inform both the research and clinical assessment of discourse change, we sought to provide detailed insights into the reliability of individual discourse tasks and grouped task sets. We believe that this approach offers valuable guidance for future research and supports clinicians in selecting discourse elicitation tasks with demonstrated psychometric robustness. The results of these tasks are presented in Supplemental Material 3. The WAB-R picture description samples were analyzed using Thematic Units. This task-specific variable is incompatible with the measures used for other discourse tasks, thereby precluding their combination. Cinderella storytelling typically yields longer and more complex language samples, and prior research has demonstrated the promising reliability of MC codes for this task. Combining it with other tasks is not necessary to achieve robust psychometric analysis.

2.8 Analysis software

All statistical analyses were performed using the psych and irr packages in RStudio 2023.12.1+402. The authors used ChatGPT (OpenAI, GPT-4, July 2025 version) to assist with RStudio coding during data processing and statistical analysis. All code outputs were carefully reviewed, tested, and modified by the authors to ensure methodological accuracy and validity.

2.8.1 Statistical analyses of intra- and inter- rater reliability

Two-way mixed intraclass correlation coefficients (ICCs) with absolute agreement were calculated for transcription-less variables and selected linguistic variables (total number of words, utterances, and CIUs).

2.8.2 Statistical analysis of test and retest reliability

Data distribution for all dependent variables was assessed using the Kolmogorov–Smirnov test at each session. As more than 70% of the variables did not meet the normality assumption, non-parametric statistical tests were used throughout the analyses. All these analyses were performed on transcription-less variables (primary

aim) and linguistic variables (secondary aim).

Although correlation is one of the most common statistical methods used to investigate test-retest reliability, the sole use of correlations in studies dealing with replicate data is insufficient, as it does not test agreement (Bland & Altman, 1986). Test-retest reliability refers to the capacity of a test or measure to replicate the same order between participants when tested twice (Kottner et al., 2011). In contrast, agreement refers to the capacity to provide the same result twice (Berchtold, 2016). Following the guidelines of Koo and Li (2016), the reliability of the test and retest sessions was evaluated using a two-way mixed ICC with absolute agreement. Also, Koo and Li's (2016) interpretation guidelines were used for all ICCs (intra-rater, inter-rater, and test–retest reliability): below .50 = poor, between .50 and .75 = moderate, between .75 and .90 = good, and above .90 = excellent. Agreement was tested using the Wilcoxon signed-rank test to evaluate whether there was a statistically significant difference between the test and the retest. We also measured the strength of the association by using Spearman's rho to assess the similarity between the test and retest. *P*-values were adjusted using the Bonferroni correction to account for multiple comparisons.

We also computed the minimal detectable change (MDC) for each dependent variable. Given the variance from the test-retest result, MDC at a 90% confidence interval (CI) (MDC90) was chosen to assess the approximate change associated with clinical change (Donoghue et al., 2009). MDC90 includes the standard error of measurement (SEM), computed using the following formula: $SEM = SD\sqrt{1-r}$, where SD is the standard deviation for the obtained score distribution and r is the correlation coefficient (i.e., ICC). The formula used to calculate MDC90 is $MDC90 = SEM * 1.65 * \sqrt{(2)}$.

Bland–Altman plots were produced to allow visual inspection of the data by examining the limits of agreement between the testing points (Altman & Bland, 1983). Bland-Altman plots are scatterplots with the Y-axis representing the difference

between the test and retest results, and the X-axis representing the mean test and retest results. The scatterplot also illustrates the limits of agreement with horizontal dashed lines at ±1.96 standard deviations of the mean of differences. A good agreement between test and retest was obtained if 95% of the data fell within these limits (Bland & Altman, 1999). These plots were created for the variables that obtained the best test-retest ICC.

2.8.3 Construct validity

Construct validity was assessed for each independent task using Spearman rho correlations between transcription-less measures (TU, MC, AC, and CoRe) and the linguistic variables of MATTR and CIU. These analyses were performed only on the test results. To control for the risk of Type I error due to multiple comparisons, *p*-values were adjusted using the Bonferroni correction.

3 Results

3.1 Rater Reliability

3.1.1 Intra-rater reliability of transcription-less variables

For the WAB-R picture description task, the intra-rater reliability of the TU was good at test and excellent at retest, and for the CoRe, it was good at test and retest. For the Cinderella storytelling task, the intra-rater reliability of the MC, AC, II, and CoRe was excellent in both the test and retest. The intra-rater reliability of the AB was excellent at test and good at retest. For the combined tasks of the Broken Window, Refused Umbrella, and Cat Rescue, the intra-rater reliability of MC and AC was good at test and excellent at retest. The CoRe intra-rater reliability was excellent in the test and retest. The intra-rater reliability results are reported in Table 6.

Table 6. Intra-rater and inter-rater reliability by task or combined tasks for each transcription-less variable.

	Intra-rater reliability	Inter-rater reliability
WAB-R picnic picture description task		
TU	Test, ICC = .899 [.324, .989] Retest, ICC = 1.000 [1.000, 1.000]	Test, ICC = 1.000 [1.000, 1.000] Retest, ICC = .989[.923, .998]
CoRe	Test, ICC = .790 [-.061, .976] Retest, ICC = .790 [-.061, .976]	Test, ICC = .709 [-.097, .953] Retest, ICC = .709 [-.097, .953]
Cinderella Story Retell		
MC	Test, ICC = .965 [.707, .996] Retest, ICC = .934 [.503, .993]	Test, ICC = .986 [.906, .998] Retest, ICC = .986 [.907, .998]
AC	Test, ICC = .990 [.904, .999] Retest, ICC = .962 [.685, .996]	Test, ICC = .971 [.808, .996] Retest, ICC = .935 [.614, .991]
AI	Test, ICC = .000 [-.811, .811] Retest, ICC = .409 [-.602, .916]	Test, ICC = .526 [-.379, .917] Retest, ICC = .342 [-.556, .872]
IC	Test, ICC = .404 [-.606, .915] Retest, ICC = .316 [-.666, .897]	Test, ICC = .192 [-.658, .827] Retest, ICC = .444 [-.466, .898]
II	Test, ICC = .992 [.926, .999]	Test, ICC = .000 [-.754, .754]

	Retest, ICC = .931 [.491, .993]	Retest, ICC = 1.000 [1.000, 1.000]
AB	Test, ICC = .921 [.432, .991]	Test, ICC = .991 [.936, .999]
	Retest, ICC = .843 [.101, .982]	Retest, ICC = .994 [.955, .999]
CoRe	Test, ICC = 1.000 [1.000, 1.000]	Test, ICC = .642 [-.218, .969]
	Retest, ICC = .987 [.880, .999]	Retest, ICC = .950 [.693, .993]
Combined tasks of Broken Window, Refused Umbrella and Cat Rescue		
MC	Test, ICC = .857 [.291, .979]	Test, ICC = .821 [.174, .973]
	Retest, ICC = .958 [.737, .994]	Retest, ICC = .965 [.772, .995]
AC	Test, ICC = .822 [.178, .973]	Test, ICC = .694 [-.127, .951]
	Retest, ICC = .978 [.855, .997]	Retest, ICC = .966 [.780, .995]
AI	Test, ICC = .670 [-.170, .946]	Test, ICC = .878 [.366, .982]
	Retest, ICC = .345 [-.553, .872]	Retest, ICC = .750 [-.010, .961]
IC	Test, ICC = .160 [-.676, .816]	Test, ICC = .189 [-.659, .826]
	Retest, ICC = .000 [-.754, .755]	Retest, ICC = .524 [-.381, .916]
II	Test, ICC = .615 [-.260, .936]	Test, ICC = .160 [-.676, .816]
	Retest, ICC = .000 [-.754, .754]	Retest, ICC = .000 [-.754, .755]
AB	Test, ICC = .727 [-.060, .957]	Test, ICC = .932 [.600, .990]

	Retest, ICC = .874 [.829, .996]	Retest, ICC = .947 [.673, .992]
CoRe	Test, ICC = .988 [.919, .998]	Test, ICC = .814 [.153, .972]
	Retest, ICC = .987 [.880, .999]	Retest, ICC = .896 [.435, .985]

Note. TU, Thematic Units; CoRe = Coherence and Reference score; MC = Main Concept total composite score; AC = Accurate and Complete; AI = Accurate and Incomplete; IC = Incorrect and Complete; II = Incorrect and Incomplete; AB = Absent. Parentheses show 95% confidence intervals (CIs) around the ICC. Koo and Li (2016) suggested the following interpretation of the intraclass correlation coefficient (ICC), including confidence intervals: below .50 = poor; between .50 and .75 = moderate; between .75 and .90 = good; and above .90 = excellent. ICC = intraclass correlation coefficient.

3.1.2 Inter-rater reliability of transcription-less variables

For the WAB-R picture description task, the inter-rater reliability of the TU was excellent in both test and retest. For the Cinderella storytelling task, the inter-rater reliability scores of the MC, AC, and AB were excellent in both the test and retest. Inter-rater reliability scores of the CoRe were moderate in the test and excellent in the retest. For the combined tasks of Broken Window, Refused Umbrella, and Cat Rescue, inter-rater reliability scores of MC were good at test and excellent at retest. The AC was moderate at test and excellent at retest. The AB and CoRe reliabilities were good in the test and retest. The results of inter-rater reliability are shown in Table 6.

3.2 Test-retest reliability of transcription-less variables

The test-retest reliability of transcription-less measures for the Picnic picture description task, the Cinderella storytelling task, and the combined tasks of the Broken Window, the Refused Umbrella, and the Cat Rescue are reported in Tables 7, 8, and 9, respectively. Given the breadth of the results, we provide a summary in the main text and refer readers to the corresponding tables for the detailed statistical values. Across tasks, the intraclass correlation coefficients (ICCs) ranged from poor to excellent. Notably, eight measures demonstrated confidence intervals that encompassed the threshold for excellent reliability. In the Picnic picture description task, TU and TU/min showed good test-retest reliability, whereas CoRe achieved excellent reliability. The Cinderella storytelling task obtained excellent reliability for MC and AB, and good reliability for AC and II. For the combined task of Broken Window, the Refused Umbrella, and the Cat Rescue, CoRe obtained excellent reliability, and MC and AC obtained good reliability. Statistical analyses demonstrated significant correlations between the sessions for most variables (11 of 19 variables). Wilcoxon analyses revealed no significant changes throughout the tasks and variables after Bonferroni correction. Test-retest reliability complete results by task are reported in Supplemental material 3.

Table 7. Test–retest reliability and descriptive statistics for the transcription-less variables of the WAB-R picture description task.

	Test		Retest		Intraclass Correlations (ICC)		Statistics		Absolute Difference	
Variable	Mean (SD)	Median [Range]	Mean (SD)	Median [Range]	ICC (95% CI)	ICC Interpretation	Spearman's ρ (p)	Wilcoxon V (p)	Mean (SD, Range)	MDC90
TU	11.65 (3.92)	13.00 [1.00–16.00]	11.96 (3.97)	13.00 [3.00–16.00]	0.791 (0.569–0.906)	Good (Moderate–Excellent)	0.733 (p = 0.001)	68.5 (p = 1.000)	1.96 (1.61, 0.00–6.00)	4.19
TU/min	8.77 (6.78)	7.35 [1.64–31.30]	8.42 (5.94)	6.86 [1.85–22.11]	0.778 (0.546–0.900)	Good (Moderate–Good)	0.756 (p = 0.000)	141.0 (p = 1.000)	2.83 (3.13, 0.02–9.26)	6.98
CoRe	6.00 (2.43)	7.00 [2.00–8.00]	6.13 (2.47)	8.00 [2.00–8.00]	0.967 (0.925–0.986)	Excellent (Excellent–Excellent)	0.887 (p = 0.000)	6.0 (p = 1.000)	0.30 (0.56, 0.00–2.00)	1.03

Note. SD = Standard Deviation; TU = Thematic Units; TU/min = Thematic Units per minute; ICC = Intraclass Correlation Coefficient; MDC90 = Minimal Detectable Change at 90% confidence intervalle. Bonferroni correction was applied to Wilcoxon and Spearman p-values.

Table 8. Descriptive statistics and summary of test–retest results for the transcription-less variables of the Cinderella storytelling task.

	Test		Retest		Intraclass Correlations (ICC)		Statistics		Absolute. Difference Mean	
Variable	Mean (SD)	Median [Range]	Mean (SD)	Median [Range]	ICC (95% CI)	ICC Interpretation	Spearman's ρ (adj. p)	Wilcoxon V (adj. p)	Mean (SD, Range)	MDC90
MC	40.25 (25.01)	37.50 [9.00–85.00]	40.65 (23.40)	38.00 [7.00–80.00]	0.910 (0.787–0.963)	Excellent (Good–Excellent)	0.903 (p = 0.000)	76.5 (p = 1.000)	7.20 (7.15, 0.00–30.00)	16.90
MC/min	11.19 (7.37)	8.63 [1.50–28.70]	10.43 (6.06)	7.72 [3.87–23.75]	0.748 (0.466–0.892)	Moderate (Poor–Good)	0.657 (p = 0.045)	147.0 (p = 1.000)	3.07 (3.69, 0.38–15.30)	7.87
AC	9.45 (8.94)	7.50 [0.00–27.00]	9.45 (8.66)	7.00 [0.00–26.00]	0.893 (0.751–0.956)	Good (Good–Excellent)	0.906 (p = 0.000)	46.0 (p = 1.000)	2.60 (3.07, 0.00–12.00)	6.69

	Test		Retest		Intraclass Correlations (ICC)		Statistics		Absolute. Difference Mean	
Variable	Mean (SD)	Median [Range]	Mean (SD)	Median [Range]	ICC (95% CI)	ICC Interpretation	Spearman's ρ (adj. p)	Wilcoxon V (adj. p)	Mean (SD, Range)	MDC90
AC/min	2.51 (2.34)	2.33 [0.00–8.87]	2.01 (1.89)	1.49 [0.00–6.71]	0.865 (0.690–0.944)	Good (Moderate–Excellent)	0.838 (p = 0.000)	95.0 (p = 1.000)	0.71 (0.98, 0.00–3.75)	1.82
AI	2.55 (1.79)	2.00 [0.00–7.00]	2.15 (1.73)	2.00 [0.00–6.00]	0.517 (0.108–0.776)	Moderate (Poor–Good)	0.560 (p = 0.215)	90.0 (p = 1.000)	1.30 (1.17, 0.00–5.00)	2.84
IC	2.25 (1.80)	2.00 [0.00–7.00]	2.80 (2.65)	3.00 [0.00–11.00]	0.256 (-0.198–0.620)	Poor (Poor–Moderate)	0.477 (p = 0.705)	46.0 (p = 1.000)	1.85 (2.08, 0.00–9.00)	4.54

	Test		Retest		Intraclass Correlations (ICC)		Statistics		Absolute. Difference Mean	
Variable	Mean (SD)	Median [Range]	Mean (SD)	Median [Range]	ICC (95% CI)	ICC Interpretation	Spearman's ρ (adj. p)	Wilcoxon V (adj. p)	Mean (SD, Range)	MDC90
II	2.30 (2.83)	1.50 [0.00–10.00]	2.40 (3.27)	1.00 [0.00–11.00]	0.803 (0.568–0.917)	Good (Moderate–Excellent)	0.630 (p = 0.061)	58.5 (p = 1.000)	1.40 (1.27, 0.00–4.00)	3.15
AB	14.50 (7.68)	13.50 [2.00–27.00]	14.10 (7.16)	14.00 [3.00–27.00]	0.921 (0.812–0.968)	Excellent (Good–Excellent)	0.894 (p = 0.000)	88.5 (p = 1.000)	2.30 (1.81, 0.00–6.00)	4.84
CoRe	4.70 (2.49)	5.00 [2.00–8.00]	4.40 (2.54)	4.00 [2.00–8.00]	0.875 (0.711–0.948)	Good (Moderate–Excellent)	0.875 (p = 0.000)	11.0 (p = 1.000)	0.50 (1.19, 0.00–5.00)	2.07

Note. SD = Standard Deviation; MC = Main Concept total composite score; AC = Accurate and Complete; AI = Accurate and Incomplete; IC = Incorrect and Complete; II = Incorrect and Incomplete; AB = Absent; ICC = Intraclass Correlation Coefficient; MDC90 = Minimal Detectable Change at 90% confidence interval. Bonferroni correction was applied to Wilcoxon and Spearman p-values.

Table 9. Test–retest reliability and descriptive statistics for the transcription-less variables of the combined tasks of the Broken Window, the Refused Umbrella, and the Cat Rescue.

	Test		Retest		Intraclass Correlations (ICC)		Statistics		Absolute mean difference	
Variable	Mean (SD)	Median [Range]	Mean (SD)	Median [Range]	ICC (95% CI)	ICC Interpretation	Spearman's ρ (p)	Wilcoxon V (p)	Mean (SD, Range)	MDC90
MC	14.09 (5.75)	15.00 [3.00–24.00]	11.57 (6.25)	10.00 [1.00–24.00]	0.799 (0.582–0.909)	Good (Moderate–Excellent)	0.755 (p = 0.001)	160.5 (p = 0.382)	3.30 (3.13, 0.00–10.00)	6.27
MC/min	17.48 (10.50)	15.17 [2.57–37.50]	13.81 (7.30)	13.04 [1.33–30.00]	0.552 (0.191–0.782)	Moderate (Poor–Good)	0.517 (p = 0.555)	192.0 (p = 1.000)	6.98 (6.02, 0.37–26.22)	14.07
AC	3.13 (1.98)	3.00 [0.00–6.00]	2.91 (2.48)	2.00 [0.00–8.00]	0.812 (0.606–0.916)	Good (Moderate–Excellent)	0.823 (p = 0.000)	90.5 (p = 1.000)	1.09 (0.85, 0.00–3.00)	2.27
AC/min	3.68 (2.64)	3.67 [0.00–8.82]	3.29 (2.70)	2.40 [0.00–8.57]	0.608 (0.270–0.813)	Moderate (Poor–Good)	0.544 (p = 0.323)	142.0 (p = 1.000)	1.92 (1.38, 0.00–4.18)	3.89

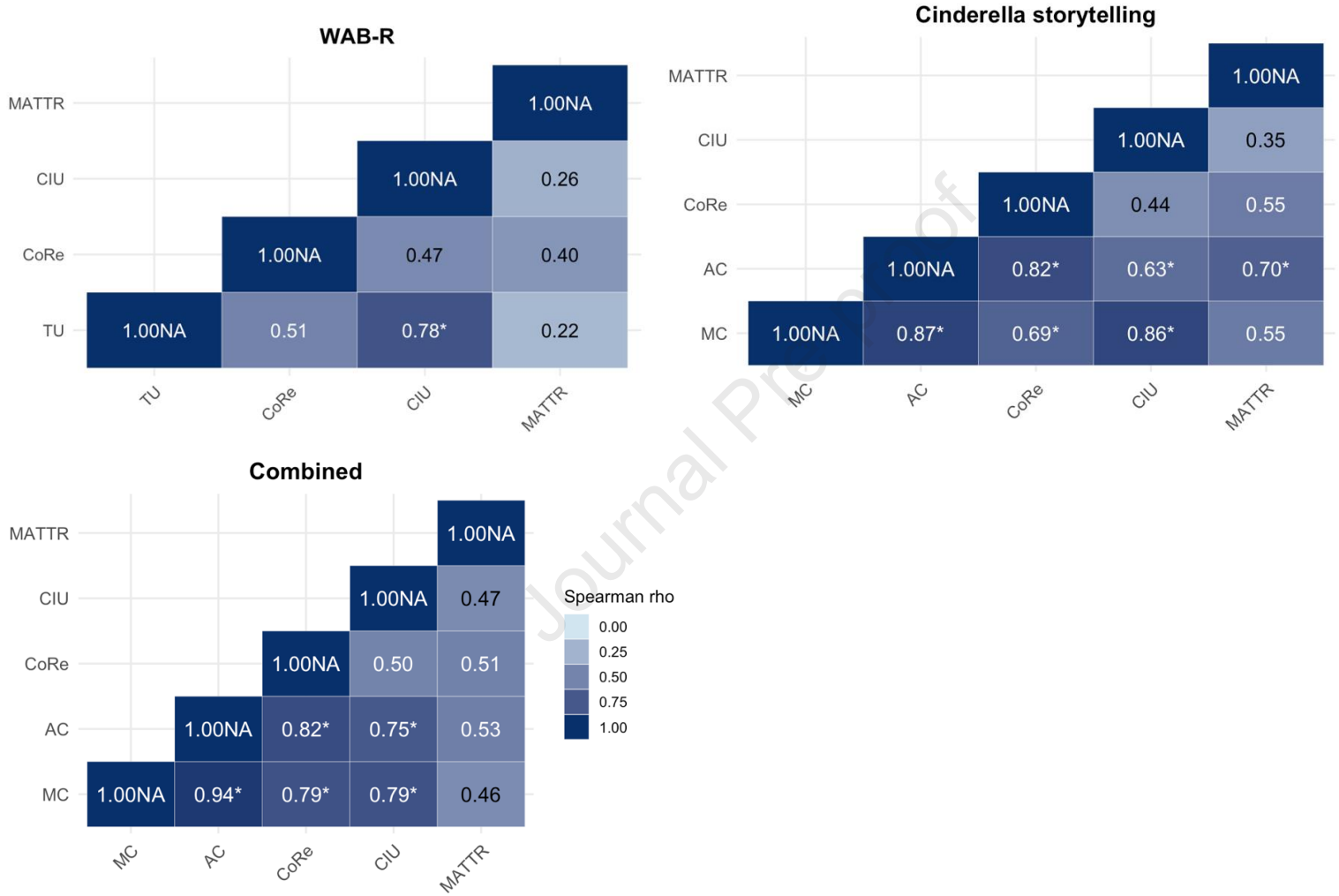
	Test		Retest		Intraclass Correlations (ICC)		Statistics		Absolute mean difference	
Variable	Mean (SD)	Median [Range]	Mean (SD)	Median [Range]	ICC (95% CI)	ICC Interpretation	Spearman's ρ (p)	Wilcoxon V (p)	Mean (SD, Range)	MDC90
AI	1.70 (0.97)	2.00 [0.00–3.00]	0.70 (1.26)	0.00 [0.00–5.00]	0.211 (-0.212–0.567)	Poor (Poor–Moderate)	0.253 (p = 1.000)	163.0 (p = 0.251)	1.43 (0.95, 0.00–3.00)	2.33
IC	0.43 (0.59)	0.00 [0.00–2.00]	0.57 (0.90)	0.00 [0.00–3.00]	-0.052 (-0.447–0.360)	Poor (Poor–Poor)	-0.026 (p = 1.000)	38.5 (p = 1.000)	0.74 (0.81, 0.00–3.00)	1.81
II	0.43 (0.66)	0.00 [0.00–2.00]	0.30 (0.56)	0.00 [0.00–2.00]	-0.005 (-0.409–0.400)	Poor (Poor–Poor)	0.108 (p = 1.000)	23.0 (p = 1.000)	0.48 (0.73, 0.00–2.00)	1.43
AB	4.26 (2.16)	4.00 [0.00–9.00]	3.22 (1.95)	3.00 [0.00–7.00]	0.673 (0.369–0.847)	Moderate (Poor–Good)	0.695 (p = 0.010)	144.0 (p = 0.430)	1.57 (1.16, 0.00–4.00)	2.74
CoRe	5.09 (2.48)	6.00 [1.00–8.00]	5.39 (2.48)	6.00 [2.00–8.00]	0.923 (0.828–0.967)	Excellent (Good–Excellent)	0.911 (p = 0.000)	22.0 (p = 1.000)	0.65 (0.78, 0.00–3.00)	1.60

Note. SD = Standard Deviation; MC = Main Concept total composite score; AC = Accurate and Complete; AI = Accurate and Incomplete; IC = Incorrect and Complete; II = Incorrect and Incomplete; AB = Absent; ICC = Intraclass Correlation Coefficient; MDC90 = Minimal Detectable Change at 90% confidence interval. Bonferroni correction was applied to Wilcoxon and Spearman p-values.

3.3 Construct validity analyses

Only statistically significant results are reported. In addition, we only reported correlations between the processing levels. For the WAB-R task, there was a strong positive correlation between TU and CIU. For the Cinderella storytelling, significant strong positive correlations were observed between MC and CIU, AC and CoRe, and AC and CIU. AC also correlated significantly with MATTR, and MC correlated significantly with CoRe. For the combined tasks of the Broken Window, Refused umbrella, and Cat Rescue, significant positive correlations were found between MC and CoRe, MC and CIU (r_s(n=15) = .786, p = .0001), AC and CoRe, and AC and CIU. The results of this task are shown in Figure 1.

Figure 1. Construct validity of transcription-less variables for the WAB-R, Cinderella storytelling, and combined tasks

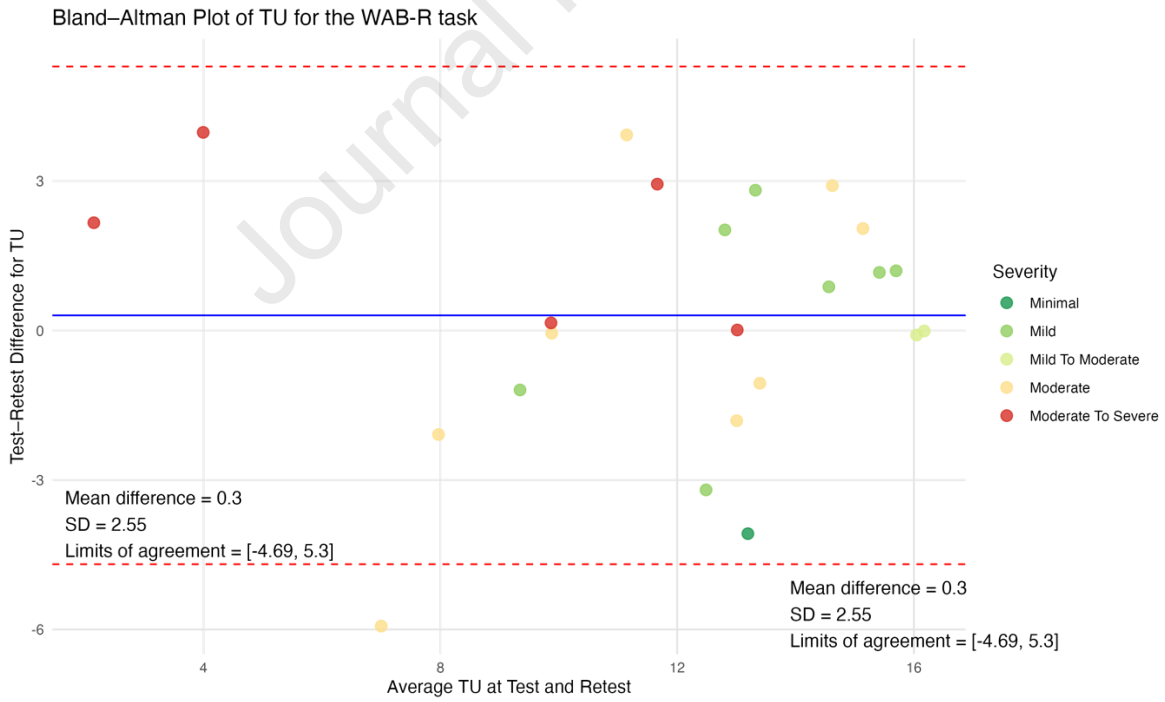


Note. WAB-R, Western Aphasia Battery-Revised picture description; TU, Thematic Units; CoRe, Coherence/reference; MC, Main Concept composite score; AC, Accurate and complete; MATTR= Moving Average Type-Token Ratio; CIU= Correct Information Units.

3.4 Agreement

Bland-Altman plots were created for the transcription-less variables of TU, MC, AC, and CoRe, as well as the linguistic variable of CIU. All of these variables obtained the best intra- and inter-rater and test-retest ICCs, ranging from good to excellent. Figure 2 illustrates the limits of agreement for the TU, CoRe, and CIU for the WAB-R task. The mean differences of agreement were close to zero for both TU and CoRe at 0.3 and 0.13. The CIU presented a mean difference of 2 between the test and retest. The TU, CoRe, and CIU demonstrated good agreement according to the standards of Bland and Altman (1999), with 95% of the data (i.e., 19 out of 20) within ±1.96 standard deviations of the mean of differences.

Figure 2. Bland-Altman plots for transcription-less variables of the WAB-R picture description task

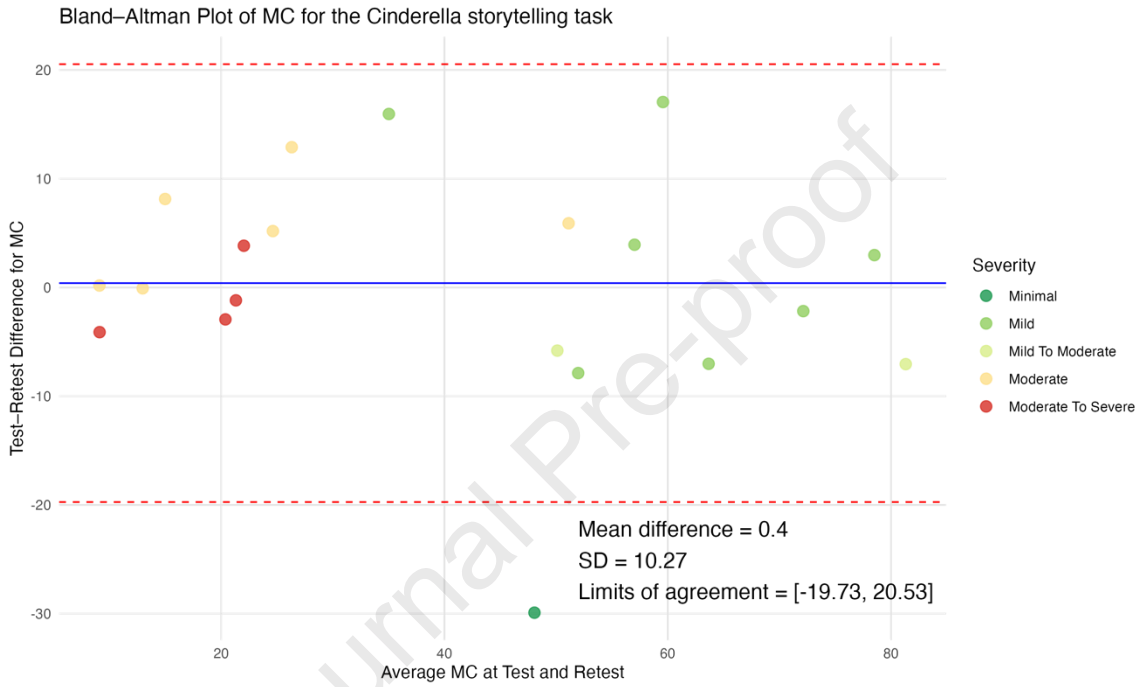


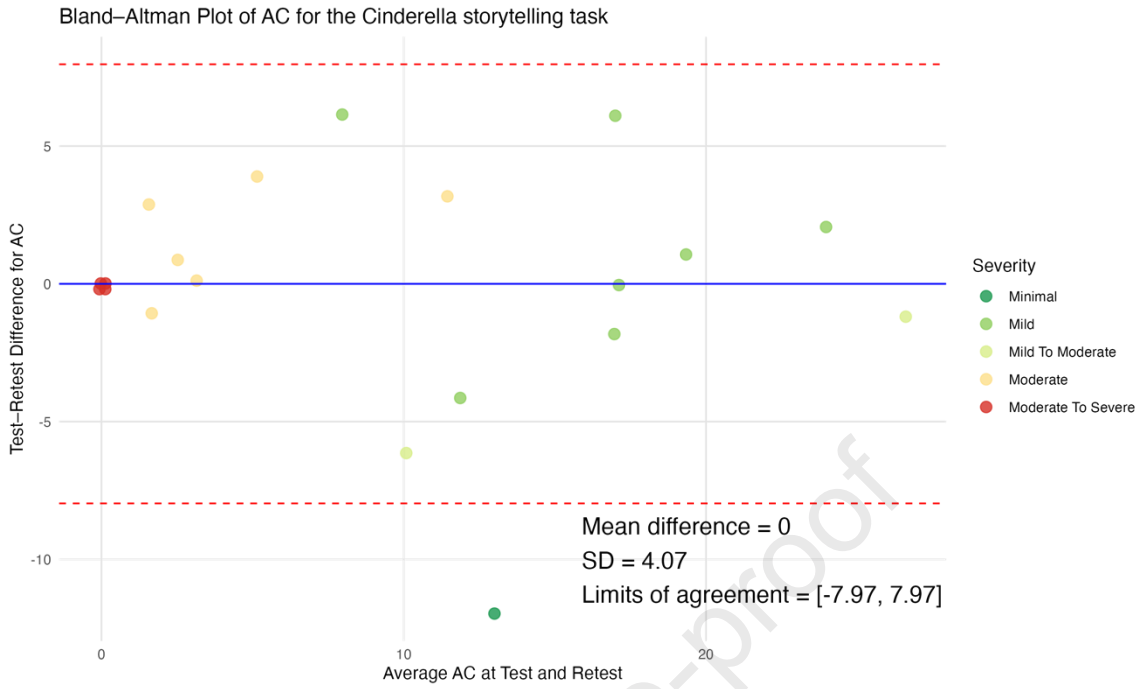
Note. CoRe, Coherence/reference; TU, Thematic Units; CIU= Correct Information Units.

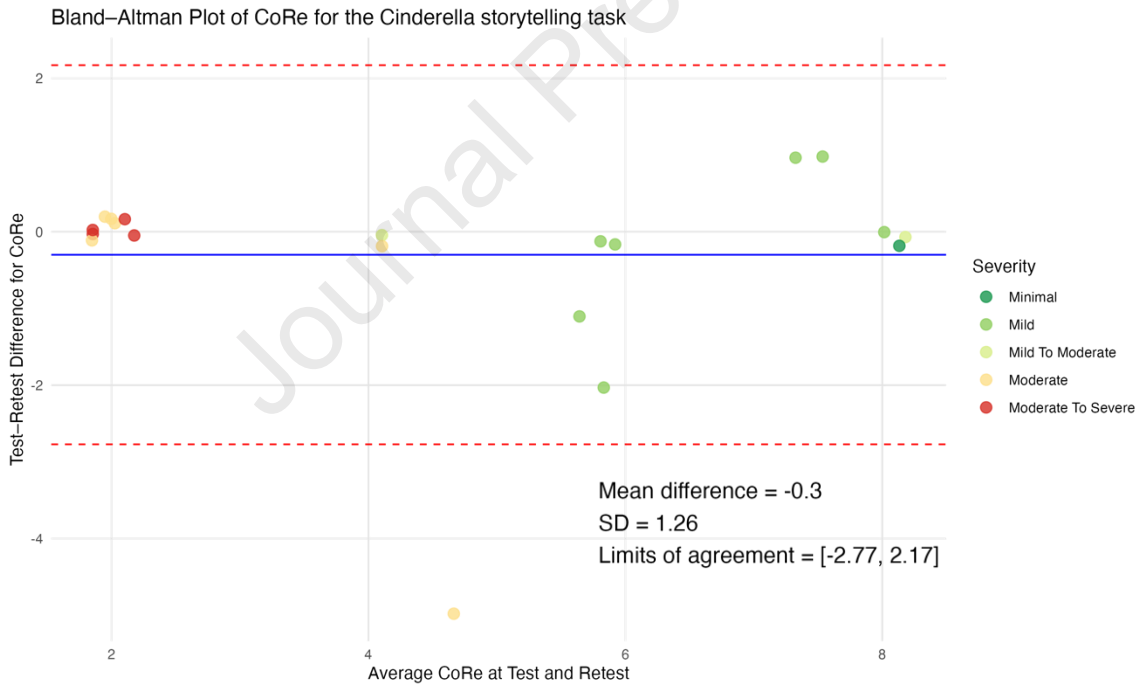
Figure 3 illustrates the limits of agreement for MC, AC, CoRe, and CIU in the Cinderella storytelling task. The mean differences of agreement were close to zero for MC, AC,

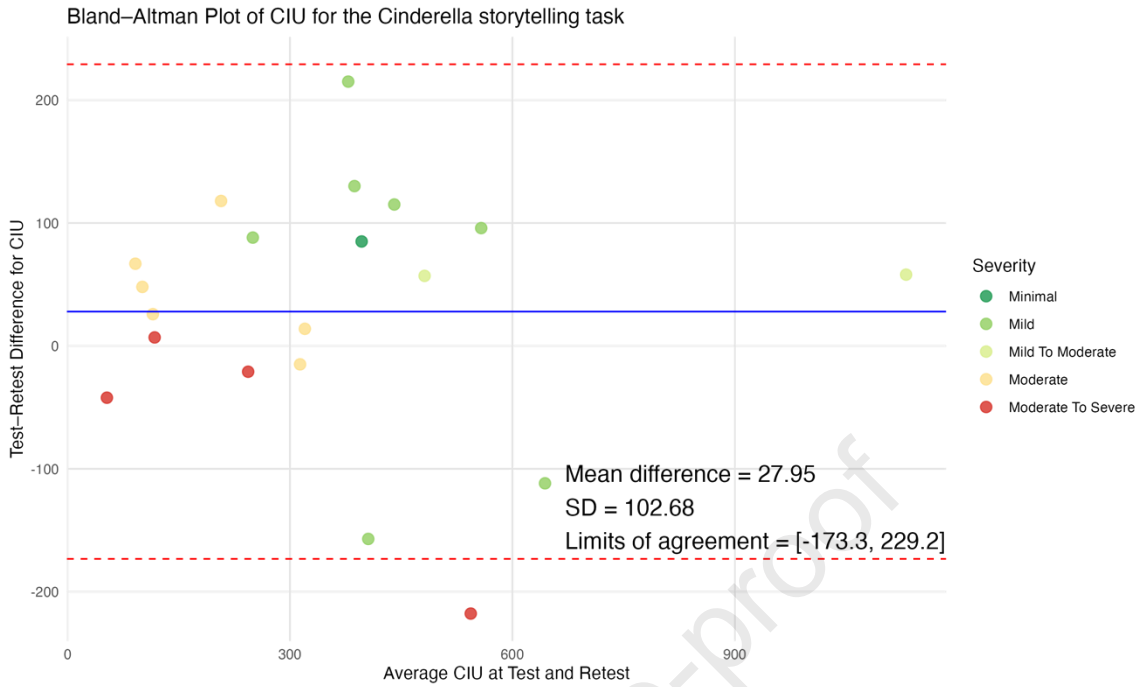
and CoRe at 0.4, 0, and -0.3. The CIU presented a mean difference of 27.95 between the test and the retest.

Figure 3. Bland-Altman plots for transcription-less variables of the Cinderella storytelling task





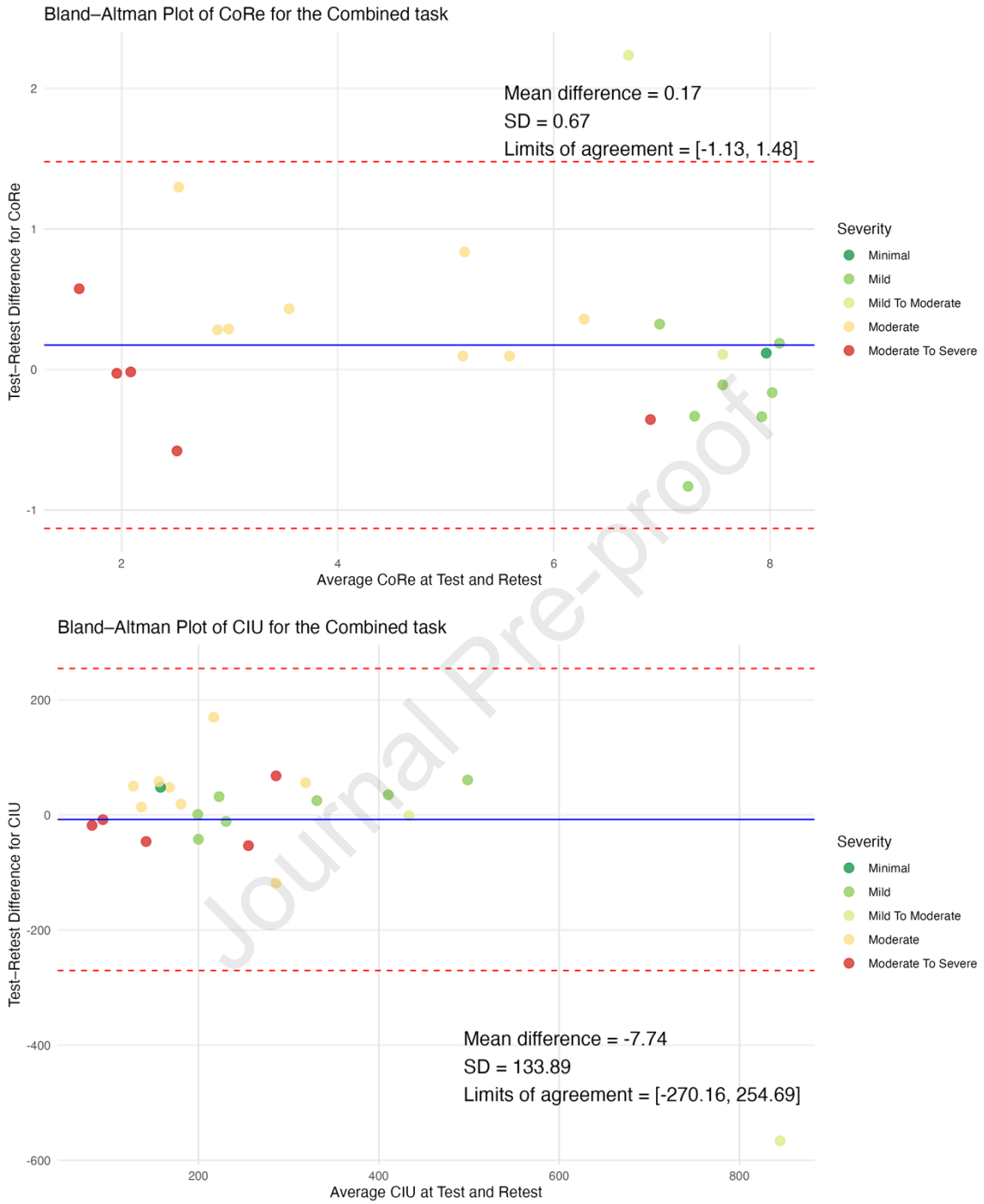




Note. CoRe, Coherence/reference; MC, Main Concept composite score; AC, Accurate and complete; CIU= Correct Information Units.

Figure 4 illustrates the limits of agreement for MC, AC, CoRe, and CIU for the combined tasks of the Broken Window, Refused umbrella, and Cat Rescue. The mean differences of agreement were close to zero for AC and CoRe at 0.91 and 0.17. MC presented a mean difference of 3.96 between test and retest, and a CIU of -7.74. For the Cinderella storytelling task and the combined tasks, MC, AC, CoRe, and CIU demonstrated good agreement according to the standards of Bland and Altman (1999), with more than 95% of the data (i.e., 22 out of 23) within ±1.96 standard deviations of the mean of differences.

Figure 4. Bland-Altman plots for transcription-less variables of the combined tasks



Note. CoRe, Coherence/reference; MC, Main Concept composite score; AC, Accurate and complete; CIU= Correct Information Units.

The Bland–Altman plots also provide a visual indication of agreement stratified by aphasia severity. Across most plots, a consistent trend emerged: participants with more severe aphasia (moderate to severe) tended to show lower mean test–retest values, whereas those with milder forms of aphasia generally demonstrated higher

means. This pattern suggests a potential relationship between aphasia severity and consistency in discourse performance. A broader examination of the WAB-R and combined task plots showed a relatively even distribution of data points across severity levels, indicating a balanced performance variability among participants. In contrast, the Cinderella storytelling task revealed a more constrained pattern in individuals with moderate-to-severe aphasia, particularly for CoRe, but also for MC and AC. In these cases, the data points were more tightly clustered, suggesting reduced discourse performance variability for this task subgroup.

3.5 Linguistic measures

3.5.1 Inter-rater reliability of transcripts

Similar to Stark et al. (2023), given the small confidence intervals and overwhelmingly excellent inter-rater reliability, reliability for the linguistic variables was computed on data across tasks (i.e., all discourse data averaged across the five tasks at Test and Retest). Inter-rater agreement was excellent (ICC >.90) for the total number of utterances, tokens, and CIU.

3.5.2 Test-retest reliability of linguistic measures

Given that the breadth of results and assessment of test-retest reliability of linguistic measures is a secondary aim, the results are not reported directly. For all tasks, 10 variables obtained excellent ICC and 16 obtained good ICC. Notably, the total words and CIUs obtained excellent ICC across stimuli. In addition, words/min obtained excellent ICC for two stimuli and good ICC for one task. The complete results of the test-retest reliability of the linguistic variables for each task independently and the combined tasks of the Broken Window, Refused umbrella, and Cat Rescue are presented in Supplemental Material 4.

3.6 Summary of the recommended variables

Our study reported the expected variability and MDC_{90} for transcription-less and linguistic discourse measures across five tasks in individuals with aphasia (PWA). Given the inherently variable nature of discourse, providing reference values for expected variability is essential to distinguish true changes over time. We provided adapted MC for the three tasks of the Broken Window, Refused Umbrella, and Cat Rescue, along with the scoring sheets in Supplemental Material 2. Table 10 summarizes the reliability statistics as well as MDC_{90} values to assist clinicians in implementing transcription-less discourse measures in their daily practice.

Table 10. Summary of recommended transcription-less variables, including MDC90. All reported ICC are good (between .75 and .90) or excellent (above .90). Only the positive significant correlations are reported.

Task	Variables	Intra-rater reliability analyses	Inter-rater reliability analyses	Test-retest reliability analyses	Construct validity analyses	MDC_{90}
WAB-R	TU	Test - good Retest - excellent	Test - Retest - excellent	Good	CIU	4.19
	CoRe	Test and Retest - good	-	Excellent	-	1.03
Cinderella	MC	Test and retest - excellent	Test and retest - excellent	Excellent	CIU and CoRe	16.90
	AC	Test and retest - excellent	Test and retest - excellent	Good	CIU, CoRe and MATTR	6.69
	CoRe	Test and retest - excellent	Test - moderate Retest - excellent	Good	MC and AC	2.07
Combined tasks of the Broken Window, the	MC	Test- good Retest - excellent	Test- good Retest - excellent	Good	AC, CoRe and CIU	6.27
	AC	Test- good	Test - moderate	Good	CoRe and CIU	2.27

Refused		Retest - excellent	Retest - excellent			
Umbrella and the Cat Rescue	CoRe	Test and retest excellent	Test and retest - good	Excellent	MC and AC	1.60

Note. This table reports the variables that obtained higher psychometric qualities across the analyses for each task. ICC, Intraclass correlation; WAB-R, Western Aphasia Battery-Revised picture description; TU, Thematic Units; CoRe, Coherence/reference; MC, Main Concept composite score; AC, Accurate and complete; II, Inaccurate and incomplete; AB= Absent, MATTR= Moving Average Type-Token Ratio; CIU= Correct Information Units. MDC90= Minimal detectable change at 90% confidence interval.

4 Discussion

This study aimed to evaluate the psychometric properties of three transcription-less discourse measures: MC, TU, and CoRe. Intra-rater, inter-rater, and test–retest reliabilities were assessed across five monologic discourse tasks in PWA. Construct validity was examined through correlations with transcript-based linguistic measures including MATTR and CIU. The MCA procedure was adapted for use in Laurentian French for three tasks: Broken Window, Refused umbrella, and Cat Rescue. In addition to these core measures, test–retest reliability was evaluated for 13 transcript-based linguistic discourse variables, and MDC_{90} was reported to support clinical interpretation.

4.1 Reliability of transcription-less measures

Briefly, our results support the clinical and research utility of transcription-less discourse measures. MC, AC, TU, and CoRe demonstrated strong psychometric properties across tasks, including good-to-excellent intra- and inter-rater reliability, robust test–retest stability, and meaningful associations with related discourse metrics. Notably, Wilcoxon signed-rank tests showed no significant differences between the test and retest scores, reinforcing the reliability of the findings. Together, these results confirm that these measures are consistent over time and across raters, thereby highlighting their potential for routine clinical monitoring and longitudinal research.

Our findings largely align with and expand previous research on MCA variables. Specifically, in line with Kong (2011), the MC variable consistently showed good-to-excellent reliability across all analyses, whether intra-rater, inter-rater, or test–retest, and across all elicitation tasks. Similarly, the Accuracy Code (AC) demonstrated good test–retest reliability across tasks and excellent rater agreement for the Cinderella storytelling task, echoing the findings of Nicholas and Brookshire (1995), Kong (2011),

and Boyle (2014). However, rater reliability for the combined task condition was somewhat more variable; intra-rater reliability was good at test and improved to excellent at retest. Similarly, inter-rater reliability increased from moderate to excellent between time points. These differences may reflect the influence of task variability on the rater performance. Notably, direct comparisons with previous studies are limited by methodological differences: Kong (2011) employed Kendall's tau, Boyle (2014) used point-to-point agreement, and neither reported rater reliability at either time point. Our findings support the use of MC and AC as reliable measures for tracking clinically meaningful changes in discourse following therapy and during natural recovery. While both MC and AC offer insights into the conveyance of essential content, the AC code may be especially appealing for routine clinical use because of its simplicity, as it can be computed directly without the need to aggregate multiple variables into a composite score, ultimately reducing the scoring burden and enhancing efficiency. In contrast to Boyle (2014), who examined the reliability across five different stimuli, our study tested the stability of MCA variables in a reduced set of tasks to enhance clinical applicability. Our findings suggest that MCA variables remain reliably measurable with a single task, supporting their use in faster assessments suited to time-limited, fatigue-prone acute care settings. As in Boyle (2014), our use of shorter samples likely contributed to the limited variability observed in inaccurate concepts. This resulted in lower reliability for these codes under the current task conditions. Confirming previous findings (Kong, 2011), the Absent (AB) code also showed excellent test–retest reliability for the Cinderella storytelling task and moderate reliability in the combined task condition. The rater agreement for AB was good to excellent across both individual and combined tasks. Taken together, our results reinforce the status of the MCA approach, particularly MC and AC variables, as psychometrically robust and clinically viable tools for discourse assessment. These findings further support the integration of MCA scoring into accessible, transcription-less discourse evaluation protocols.

Regarding the TU captured for the WAB-R picture description task, this measure obtained excellent intra- and inter-rater reliability, expanding previous studies of inter-rater reliability (Brisebois et al., 2020, 2022). As expected, test-retest reliability was better in PWA than in controls (Marcotte et al., 2024). More precisely, the reliability was excellent, supporting the use of this measure in future research and assessing clinical changes. Furthermore, similar to MCA variables, TU measurement requires minimal administration time, making it ideal for acute-care aphasia assessment.

Among the variables examined, CoRe was the only transcription-less measure evaluated across all tasks, with rater reliability varying depending on the specific discourse task. For the WAB-R picture description task, the intra- and inter-rater reliabilities were consistently moderate. In contrast, the intra-rater reliability was excellent for the Cinderella storytelling and combined tasks at both timepoints. Inter-rater reliability was moderate at test and excellent at retest for the Cinderella storytelling task and consistently good for the combined tasks. This variability aligns with the nature of the construct being scored; coherence and referential clarity may be more difficult to reliably assess in tasks that lack a structured narrative. For instance, in picture descriptions, coherence may play a less central role in communicative success, as the primary demand is to describe visual elements rather than organize events into a logical sequence. Notably, intra-rater reliability was excellent for the two tasks—a critical strength, especially in clinical contexts where the same clinician may be responsible for tracking changes over time. Ulatowska et al. (2003) reported a 90% point-to-point inter-rater agreement for coherence and 70% for reference; our findings expand this work by evaluating intra-rater reliability. The CoRe also demonstrated excellent test–retest reliability across all stimuli, including individual tasks contributing to the combined score, confirming its stability over time. From a practical perspective, our team found that scoring CoRe required concentration even after training but was nonetheless considered straightforward. The only challenge arose with one participant, whose sparse output made it difficult

to judge coherence on such a small sample, which may account for the observed lower inter-rater reliability. Overall, the raters reported that the scoring categories were intuitive and easy to apply. In summary, the CoRe appears to be a promising measure for capturing clinically meaningful changes in discourse coherence and referential appropriateness.

In summary, this study provides strong evidence that transcription-less discourse measures, including MC, AC, TU, and CoRe, are reliable and clinically viable tools for assessing aphasia. These variables showed consistent performance across tasks and rating conditions, with good-to-excellent inter- and intra-rater agreement, robust test–retest reliability, and meaningful construct validity. Crucially, each measure taps into a distinct level of discourse processing, offering a multidimensional view of communication that aligns with how PWA use language in daily life. Given the central role of discourse in social participation and recovery and the longstanding need for efficient and scalable assessment tools, these findings represent a meaningful advance. Our results support the integration of these measures into routine clinical workflows, particularly in time-constrained settings, and highlight their potential for monitoring progress, guiding treatment, and improving the ecological validity of aphasia assessments. Table 10 summarizes the psychometric results for the variables that showed the most robust quality.

4.2 Refining discourse assessment: the role of task selection

Our findings emphasize the critical role of task selection when using discourse measures to assess aphasia severity or track changes over time. As shown in previous work (Stark et al., 2023), aphasia severity can affect discourse performance differently across task types. The Bland–Altman plots in our study visually illustrated this pattern: participants with moderate to severe aphasia generally produced lower mean values at both time points, while those with milder aphasia demonstrated higher and more stable scores. However, the relationship between severity and discourse output was

inconsistent across all tasks.

The WAB-R picture description and combined task condition revealed relatively even data dispersion across severity levels, suggesting that these tasks allow for more individualized performance and are better suited to capturing variability in language profiles. This aligns with the known variability in discourse production, even among neurotypical speakers (Marcotte et al., 2024), and reinforces the need to assess discourse across multiple elicitation contexts. In contrast, the Cinderella storytelling task yielded more constrained scores in participants with greater impairments, particularly for CoRe, MC, and AC, suggesting that it may be less accessible to individuals with moderate to severe aphasia. However, this same task showed high sensitivity to subtle lexical-semantic changes in individuals with mild cognitive impairment and latent aphasia (Stark et al., 2025), underscoring its value in detecting more nuanced deficits in less affected populations.

Overall, tasks that include visual support (e.g., WAB-R, Cat Rescue) or structured event sequences (e.g., Broken Windows and Refused umbrellas) appear to offer a more balanced challenge, accommodating a wider range of severity. These tasks may better capture the heterogeneous nature of aphasia by eliciting representative performances across diverse profiles.

Our findings support the need for psychometric validation not only at the group level, but also across severity subgroups. Tailoring reliability assessments, for instance, through double baselines or single-subject designs, can increase the precision and clinical utility of discourse outcome measures. Ultimately, assessing performance across multiple discourse tasks remains essential to accurately characterize aphasia and capture meaningful changes over time.

4.3 Construct validity and selection of outcome measures

To assess construct validity, we examined the relationships among transcription-less

variables and between these variables and transcript-based linguistic measures (CIU, MATTR), guided by the LUNA framework. Each transcription-less variable targets specific discourse processing levels: MC and AC reflect propositional organization; TU, macrostructural planning; and CoRe, macrostructural and pragmatic levels. Lexical informativeness (CIU) and lexical diversity (MATTR) represent linguistic formulations. While all transcription-less measures rely on language, their sensitivity varies based on their theoretical targets and task demands.

Significant positive correlations across variables support their construct validity and align with LUNA's multilevel model. For instance, TU significantly correlated with CIU during the WAB-R picture description task, suggesting that thematic relevance (macrostructure) aligns with lexical informativeness (linguistic level), which is consistent with Brisebois et al. (2022). The lack of correlation between TU and CoRe in this task highlights the influence of task structure. In description-based tasks, coherence and referential clarity may play a lesser role in communicative success.

In contrast, MC and AC showed strong correlations with both CIU and CoRe in the Cinderella storytelling and combined tasks, indicating that the ability to convey accurate and complete content draws on propositional, linguistic, and pragmatic processing. These findings mirror those of Linnik et al. (2022), who demonstrated that both micro- and macrostructural features shape coherence, and Wright and Capilouto (2012), who found that lexical informativeness and diversity contribute to global coherence. AC's correlation with MATTR further suggests a relationship between precise conceptual encoding and lexical richness, echoing Kong's (2011) findings. Collectively, these results reinforce that transcription-less variables reflect interactions across discourse levels in clinically meaningful ways.

In addition to construct validity, our study contributes to Boyle's (2020) framework for selecting discourse outcome measures, which emphasizes matching variables to client needs, the clinical context, and psychometric quality. We address the latter by

providing robust evidence for two key psychometric properties: scoring reliability and test–retest stability. While responsiveness to change was beyond the scope of this study, prior work suggests that transcription-less variables are sensitive to clinical changes. For example, TU has captured recovery between acute and chronic stages (Brisebois et al., 2020, 2022), while MC has been used to differentiate between PWA and various comparison groups, including individuals with subclinical aphasia (Fromm et al., 2017) and primary progressive aphasia (Dalton et al., 2020). In Laurentian French, MC demonstrated treatment sensitivity following a phonological component analysis (Masson-Trottier et al., 2022). CoRe, while less studied in intervention contexts, has shown meaningful correlations with the WAB-AQ and discourse coherence ratings (Ulatowska et al., 2003), pointing to its potential as a treatment outcome measure.

Although CIU and MATTR are transcript-based, our study supports their use in Laurentian French. The CIU is well established in English-language studies (Boyle et al., 2022; Peach & Reuter, 2010), and it is a validated measure of lexical diversity (Rose et al., 2016). As Dipper et al. (2020) noted, over 500 outcome measures have been used in aphasia discourse intervention studies, mostly at the word level, with no clear consensus on best practices. Our findings support MC, TU, and CoRe as practical and reliable alternatives that go beyond lexical metrics and offer insights into multilevel discourse processing.

4.4 Limitations and future directions

This study had several limitations. First, we focused on structured tasks—picture description, storytelling, and sequential storytelling—which do not capture more spontaneous or personal narratives that are often central to aphasia assessment. Recent findings suggest that unstructured conversations can be assessed reliably and reveal features that are not present in structured tasks (Leaman & Edmonds, 2021). While differences exist between genres, key discourse elements, such as

communicative success, subject-verb use, and complete utterances, are strongly correlated across tasks. Moreover, structured tasks, such as picture descriptions, have been shown to predict performance in spontaneous speech, particularly regarding word count and correct information units (Doyle et al., 1995). Second, our measures did not target the non-verbal or paralinguistic dimensions of communication, which are underrepresented in current transcription-less tools (Stark & Dalton, 2024). Given the growing evidence on the importance of multimodal communication in aphasia (Dutta & Mohapatra, 2024), future research should integrate these aspects to capture a more comprehensive picture of discourse competence. Third, while our study assessed intra- and inter-rater reliability, as in Stark et al. (2023) and Kong (2011), we included raters with varying experience levels: an experienced speech-language pathologist, novice clinician, and second-year student in a professional SLP master program. Although the rating team included one novice and one SLP student, good-to-excellent reliability was achieved for at least two variables across all stimuli. This suggests that the scoring procedures may be robust to some variation in rater experience. However, future studies should directly examine the influence of rater expertise on scoring consistency. For instance, Casilio et al. (2019) found that raters' experiences affect the reliability of particular speech and language features but not others. Future studies should also examine the influence of demographic and clinical variables, including aphasia severity, age, education, sex, and bilingualism, on discourse measure reliability, for instance through partial correlation analyses. Our sample size precluded such analyses in the present study. Notably, 16 out of 23 participants were bilingual or multilingual, which reflects the composition of our clinical population and enhances the ecological validity of our findings, but also highlights the need for future work specifically examining the role of multilingualism in discourse reliability. Finally, while MC, TU, and CoRe offer practical means of assessing discourse in diverse clinical contexts, they should be used with a comprehensive aphasia battery that includes constrained language tasks. This

combined approach ensures a more complete assessment of the language and communication abilities of PWA.

5 Conclusion

This study provides the first psychometric validation of transcription-less discourse measures—Main Concept Analysis (MCA), Thematic Units (TU), and coherence/reference (CoRe)—across five discourse tasks in Laurentian French speakers with aphasia. All three measures demonstrated strong reliability and feasibility in assessing discourse in individuals with chronic aphasia. CoRe has emerged as particularly robust across tasks and rating conditions, reinforcing its value for clinical and research applications.

Importantly, these findings underscore the potential of culturally adapted, transcription-less measures to track discourse level changes over time. They also highlighted the impact of task selection as different elicitation contexts captured distinct discourse profiles and varied in their sensitivity across severity levels. By offering efficient, scalable, and reliable alternatives to transcript-based methods, this study lays the groundwork for integrating these measures and their associated stimuli into routine clinical practice, ultimately improving the accessibility and ecological validity of discourse assessment in aphasia.

CRediT authorship contribution statement

- Amélie Brisebois: Conceptualization, Writing – review & editing, Writing – original draft, Methodology, Software, Formal analysis, Data curation.
- Simona Maria Brambati: Conceptualization, Writing – review & editing, Funding acquisition, Supervision.
- Véronique Desjardins: Formal analysis, Writing – review & editing.
- Éva Marois: Formal analysis, Writing – review & editing.
- Julie Bélanger: Formal analysis, Writing – review & editing.

- Karine Marcotte: Supervision, Conceptualization, Writing – review and editing, Writing – original draft Conceptualization, Funding acquisition, Resources, Data curation.

6 Supplemental Materials

Supplemental Material 1. Best Practice Guidelines for Reporting Spoken Discourse in Aphasia and Neurogenic Communication Disorders.

Supplemental Material 2. Scoring templates.

Supplemental Material 3. Test-retest reliability of transcription-less variables for individual tasks.

Supplemental Material 4. Test-retest reliability of linguistic variables.

Acknowledgments

The authors express their warm gratitude to the participants and their families who gave their time and expertise to our project. We also thank the speech-language pathologists who contributed to the project: Marie-Hélène Lavoie and Anne-Claire Albisetti for their help with participant recruitment, and Marie-Michelle Brouillard for her assistance with scoring.

7 Data Availability Statement

We obtained ethical approval to share individual raw data (e.g., audio and language sample transcriptions) for the participants who consented to it. Our research team is currently preparing these materials for dissemination in AphasiaBank database https://talkbank.org/aphasia/.

Funding statement

K.M. holds a Career Award from the "Fonds de Recherche du Québec – Santé" (https://doi.org/10.69777/330547). A.B. holds a scholarship from the "Fonds de Recherche du Québec – Santé" (https://doi.org/10.69777/287802). This work was supported by a Heart and Stroke Foundation of Canada grant (G-23-0034174).

Use of Generative Artificial Intelligence Tools

The authors used ChatGPT (OpenAI, GPT-4, July 2025 version) to support RStudio coding and improve the clarity of language during manuscript preparation. All outputs were reviewed and edited to ensure accuracy, originality, and compliance with ethical standards.

Ethics approval statement

This project is part of a larger study approved by the ethics review board of the Center intégré universitaire de santé et de services sociaux du Nord-de-l'Ile-de Montréal (CIUSSS-NÎM; #2020-1900) which sought to investigate longitudinal discourse changes following a stroke.

Patient consent statement

All participants provided written informed consent to participate in the present study.

Permission to reproduce material from other sources

n/a

References

Altman, D. G., & Bland, J. M. (1983). Measurement in Medicine: The Analysis of Method Comparison Studies. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *32*(3), 307–317. https://doi.org/10.2307/2987937

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 & Communication Disorders*, *60*(3), e70043. https://doi.org/10.1111/1460-6984.70043

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

Andreetta, S., & Marini, A. (2015). The effect of lexical deficits on narrative disturbances in fluent aphasia. *Aphasiology*, *29*(6), 705–723. https://doi.org/10.1080/02687038.2014.979394

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

Berchtold, A. (2016). Test–retest: Agreement or reliability? *Methodological Innovations*, *9*. https://doi.org/10.1177/2059799116672875

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet (London, England)*, *1*(8476), 307–310.

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research, 8*(2), 135–160. https://doi.org/10.1177/096228029900800204open_in_new

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

Boyle, M. (2020). Choosing Discourse Outcome Measures to Assess Clinical Change. *Seminars in Speech and Language*, *41*(1), 1–9. https://doi.org/10.1055/S-0039-3401029

Brisebois, A., Brambati, S. M., Boucher, J., Rochon, E., Leonard, C., Désilets-Barnabé, M., Desautels, A., & Marcotte, K. (2022). A longitudinal study of narrative discourse in post-stroke aphasia. *Aphasiology*, *36*(7), 805–830. https://doi.org/10.1080/02687038.2021.1907295

Brisebois, A., Brambati, S. M., Désilets-Barnabé, M., Boucher, J., García, A. O., Rochon, E., Leonard, C., Desautels, A., & Marcotte, K. (2020). The importance of thematic informativeness in narrative discourse recovery in acute post-stroke aphasia. *Aphasiology*, *34*(4), 472–491. https://doi.org/10.1080/02687038.2019.1705661

Brisebois, A., Brambati, S. M., Jutras, C., Rochon, E., Leonard, C., Zumbansen, A., Anglade, C., & Marcotte, K. (2023). Adaptation and Reliability of the Cinderella Story Retell Task in Canadian French Persons Without Brain Injury. *American Journal of Speech-Language Pathology*, *32*, 2871–2888. https://doi.org/10.1044/2023_AJSLP-23-00101

Brisebois, A., Brambati, S. M., Rochon, E., Leonard, C., & Marcotte, K. (2023). The longitudinal trajectory of discourse from the hyperacute to the chronic phase

in mild to moderate poststroke aphasia recovery: A case series study. *International Journal of Language & Communication Disorders*, *58*(4), 1061–1081. https://doi.org/10.1111/1460-6984.12844

Brookshire, R. H., & Nicholas, L. E. (1994a). 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.

Brookshire, R. H., & Nicholas, L. E. (1994b). *Test-Retest Stability of Measures of Connected Speech in Aphasia* [Clinical Aphasiology Paper]. Clinical Aphasiology; Pro-Ed. http://aphasiology.pitt.edu/163/

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

Casilio, M., Rising, K., Beeson, P. M., Bunton, K., & Wilson, S. M. (2019). Auditory-Perceptual Rating of Connected Speech in Aphasia. *American Journal of Speech-Language Pathology*, *28*(2), 550–568. https://doi.org/10.1044/2018_AJSLP-18-0192

Colin, C., & Le Meur, C. (2016). *Adaptation du projet aphasiabank à la langue française: Contribution pour une évaluation informatisée du discours oral de patients aphasiques: Vol. Mémoire présenté en vue de l'obtention du Certificat de Capacité d'Orthophonie*. Université Paul Sabatier, Toulouse III. Toulouse, France.

Deng, B.-M., Liang, L.-S., Zhao, J.-X., Zheng, H.-Q., & Hu, X.-Q. (2024). Correct Information Unit Analysis in Different Discourse Tasks Among Persons With Anomic Aphasia Based on Mandarin AphasiaBank. *American Journal of Speech-Language Pathology*, *33*(2), 800–813. https://doi.org/10.1044/2023_AJSLP-23-00217

Dipper, L., Marshall, J., Boyle, M., Botting, N., Hersh, D., Pritchard, M., & Cruice, M. (2020). Treatment for improving discourse in aphasia: A systematic review and synthesis of the evidence base. *Aphasiology*. https://doi.org/10.1080/02687038.2020.1765305

Dipper, L., Marshall, J., Boyle, M., Hersh, D., Botting, N., & Cruice, M. (2021a). Creating a Theoretical Framework to Underpin Discourse Assessment and Intervention in Aphasia. *Brain Sciences*, *11*(2), 183. https://doi.org/10.3390/brainsci11020183

Dipper, L., Marshall, J., Boyle, M., Hersh, D., Botting, N., & Cruice, M. (2021b). Creating a Theoretical Framework to Underpin Discourse Assessment and Intervention in Aphasia. *Brain Sciences*, *11*(2), Article 2. https://doi.org/10.3390/brainsci11020183

Dipper, L., & Pritchard, M. (2017). Discourse: Assessment and Therapy. In *Advances in Speech-language Pathology*. InTech. https://doi.org/10.5772/intechopen.69894

Donoghue, D., Physiotherapy Research and Older People (PROP) group, & Stokes, E. K. (2009). How much change is true change? The minimum detectable change of

the Berg Balance Scale in elderly people. *Journal of Rehabilitation Medicine*, *41*(5), 343–346. https://doi.org/10.2340/16501977-0337

Doyle, P. J., Goda, A. J., & Spencer, K. A. (1995). The Communicative Informativeness and Efficiency of Connected Discourse by Adults With Aphasia Under Structured and Conversational Sampling Conditions. *American Journal of Speech-Language Pathology*, *4*(4), 130. https://doi.org/10.1044/1058-0360.0404.130

Dutta, M., & Mohapatra, B. (2024). Expanding the scope: Multimodal dimensions in aphasia discourse analysis—preliminary findings. *Frontiers in Human Neuroscience*, *18*. https://doi.org/10.3389/fnhum.2024.1419311

Edmonds, L. A., & Babb, M. (2011). Effect of Verb Network Strengthening Treatment in Moderate-to-Severe Aphasia. *American Journal of Speech-Language Pathology*, *20*(2), 131–145. https://doi.org/10.1044/1058-0360(2011/10-0036)

Fitzpatrick, R., Davey, C., Buxton, M. J., & Jones, D. R. (1998). Evaluating patient-based outcome measures for use in clinical trials: A review. *Health Technology Assessment*, *2*(14). https://doi.org/10.3310/hta2140

Frederiksen, C. H., & Stemmer, B. (1993). Conceptual processing of discourse by a right hemisphere brain-damaged patient. In H. Brownell & Y. Joanette (Ed.), *Narrative discourse in neurologically impaired and normal aging adults* (pp. 239–278). Singular Publishing Group.

Fromm, D., Greenhouse, J., Hou, K., Russell, G. A., Cai, X., Forbes, M., Holland, A., & MacWhinney, B. (2016). Automated Proposition Density Analysis for Discourse in Aphasia. *J Speech Lang Hear Res*, *59*(5), 1123–1132. https://doi.org/10.1044/2016_jslhr-l-15-0401

Goodglass, H., Kaplan, E., Barresi, B., Goodglass, H., Goodglass, H., Goodglass, H., Goodglass, H., Goodglass, H., Kaplan, E., & Kaplan, E. (2001). *The Boston Diagnostic Aphasia Examination: BDAE-3 long form kit*.

Greenslade, K. J., Stuart, J. E. B., Richardson, J. D., Dalton, S. G. H., & Ramage, A. E. (2020). Macrostructural analyses of cinderella narratives in a large nonclinical sample. *American Journal of Speech-Language Pathology*, *29*(4), 1923–1936. https://doi.org/10.1044/2020_AJSLP-19-00151

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., & Duda, S. N. (2019). The REDCap consortium: Building an international community of software platform partners. *Journal of Biomedical Informatics*, *95*, 103208. https://doi.org/10.1016/j.jbi.2019.103208

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

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

Kong, A. P.-H. (2011). The main concept analysis in cantonese aphasic oral discourse: External validation and monitoring chronic aphasia. *Journal of Speech, Language, and Hearing Research*, *54*(1), 148–159. https://doi.org/10.1044/1092-4388(2010/09-0240)

Kong, A. P.-H., Cheung, C. Y.-N., & Wong, C. W.-Y. (2025). Establishing Norm of Connected Speech Measures for Descriptive Discourses in Cantonese-Speaking Adults. *International Journal of Language & Communication Disorders*, *60*(3), e70055. https://doi.org/10.1111/1460-6984.70055

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

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/J.JCM.2016.02.012

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, *64*(1), 96–106. https://doi.org/10.1016/j.jclinepi.2010.03.002

Leaman, M. C., & Edmonds, L. A. (2021). *Assessing Language in Unstructured Conversation in People With Aphasia: Methods, Psychometric Integrity, Normative Data, and Comparison to a Structured Narrative Task*. 1–22. https://doi.org/10.1044/2021_JSLHR-20-00641

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–75. https://doi.org/10.1017/S0140525X99001776

Linnik, A., Bastiaanse, R., Stede, M., & Khudyakova, M. (2022). Linguistic mechanisms of coherence in aphasic and non-aphasic discourse. *Aphasiology*, *36*(2), 123–146. https://doi.org/10.1080/02687038.2020.1852527

Macoir, J., Gauthier, C., & Jean, C. (2015). *Batterie d'Évaluation Cognitive du Langage chez l'Adulte (BECLA)* (U. Laval, Ed.).

Macoir, J., Chagnon, A., Hudon, C., Lavoie, M., & Wilson, M. A. (2021). TDQ-30-A New Color Picture-Naming Test for the Diagnostic of Mild Anomia : Validation and Normative Data in Quebec French Adults and Elderly. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *36*(2), 267-280. https://doi.org/10.1093/arclin/acz048

MacWhinney, B. (2000). *The CHILDES Project: Tolls for Analzying Talk: Vol. 3rd Editio*. Lawrence Erlbaum Associates.

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

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

Marcotte, K., Roy, A., Brisebois, A., Jutras, C., Leonard, C., Rochon, E., & Brambati, S. M. (2024). Reliability of the picture description task of the Western Aphasia Battery – revised in Laurentian French persons without brain injury. *The Clinical Neuropsychologist*, *0*(0), 1–29. https://doi.org/10.1080/13854046.2024.2340777

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

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

Nespoulous, J. L., Lecours, A. R., Lafond, D., Lemay, M. A., Puel, M., Joanette, Y., Cot, F., & Rascol, A. (1992). *Protocole Montréal-Toulouse d'examen linguistique de l'aphasie: MT-86 module standard initial, M1b(2e édition révisée par Renée Béland et Francine Giroux)*. Ortho Edition.

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

Nicholas, L. E., & Brookshire, R. H. (1995). Presence, completeness, and accuracy of main concepts in the connected speech of non-brain-damaged adults and adults with aphasia. *Journal of Speech and Hearing Research*, *38*(1), 145–156. https://doi.org/10.1044/jshr.3801.145

Prins, R., & Bastiaanse, R. (2004). Review. *Aphasiology*, *18*(12), 1075–1091. https://doi.org/10.1080/02687030444000534

Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. (2018). Psychometric properties of discourse measures in aphasia: Acceptability, reliability, and validity. *International Journal of Language & Communication Disorders*, *53*(6), 1078–1093. https://doi.org/10.1111/1460-6984.12420

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

Sherratt, S. (2007). Multi-level discourse analysis: A feasible approach. *Aphasiology*, *21*(3–4), 375–393. https://doi.org/10.1080/02687030600911435

Sherratt, S., & Bryan, K. (2019). Textual cohesion in oral narrative and procedural discourse: The effects of ageing and cognitive skills. *International Journal of*

Language & Communication Disorders, *54*(1), 95–109. https://doi.org/10.1111/1460-6984.12434

Sloetjes, H., & Wittenburg, P. (2008). Annotation by category-ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. https://archive.mpi.nl/tla/elan

Stark, B. C., Alexander, J. M., Hittson, A., Doub, A., Igleheart, M., Streander, T., & Jewell, E. (2023). Test-Retest Reliability of Microlinguistic Information Derived From Spoken Discourse in Persons With Chronic Aphasia. *Journal of Speech, Language, and Hearing Research: JSLHR*, *66*(7), 2316–2345. https://doi.org/10.1044/2023_JSLHR-22-00266

Stark, B. C., Bryant, L., Themistocleous, C., den Ouden, D.-B., & Roberts, A. C. (2022). Best practice guidelines for reporting spoken discourse in aphasia and neurogenic communication disorders. *Aphasiology*, *0*(0), 1–24. https://doi.org/10.1080/02687038.2022.2039372

Stark, B. C., & Dalton, S. G. (2024). A scoping review of transcription-less practices for analysis of aphasic discourse and implications for future research. *International Journal of Language & Communication Disorders*, *59*(5), 1734–1762. https://doi.org/10.1111/1460-6984.13028

Stark, B. C., Dutta, M., Murray, L. L., Fromm, D., Bryant, L., Harmon, T. G., Ramage, A. E., & Roberts, A. C. (2021). Spoken discourse assessment and analysis in aphasia: An international survey of current practices. *Journal of Speech, Language, and*

Hearing Research, *64*(11), 4366–4389. https://doi.org/10.1044/2021_JSLHR-20-00708

Ulatowska, H. K., Olness, G. S., Wertz, R. T., Samson, A. M., Keebler, M. W., & Goins, K. E. (2003). Relationship between discourse and Western Aphasia Battery performance in African Americans with aphasia. *Aphasiology*. https://doi.org/10.1080/0268703034400102

Wallace, S. J., Worrall, L., Rose, T., Le Dorze, G., Cruice, M., Isaksen, J., Kong, A. P.-H., Simmons-Mackie, N., Scarinci, N., & Gauvreau, C. A. (2017). Which outcomes are most important to people with aphasia and their families? An international nominal group technique study framed within the ICF. *Disability and Rehabilitation*, *39*(14), 1364–1379. https://doi.org/10.1080/09638288.2016.1194899

Wallace, S. J., Isaacs, M., Ali, M., & Brady, M. C. (2023). Establishing reporting standards for participant characteristics in post-stroke aphasia research : An international e-Delphi exercise and consensus meeting. *Clinical Rehabilitation*, *37*(2), 199-214. https://doi.org/10.1177/02692155221131241

Whitworth, A., Leitão, S., Cartwright, J., Webster, J., Hankey, G. J., Zach, J., Howard, D., & Wolz, V. (2015). NARNIA: a new twist to an old tale. A pilot RCT to evaluate a multilevel approach to improving discourse in aphasia. *Aphasiology*, *29*(11), 1345–1382. https://doi.org/10.1080/02687038.2015.1081143

Wright, H. H., & Capilouto, G. J. (2012). Considering a multi-level approach to understanding maintenance of global coherence in adults with aphasia. *Aphasiology*, *26*(5), 656–672. https://doi.org/10.1080/02687038.2012.676855

Zhang, M., Geng, L., Yang, Y., & Ding, H. (2020). Cohesion in the discourse of people with post-stroke aphasia. *Https://Doi.Org/10.1080/02699206.2020.1734864*, *35*(1), 2–18. https://doi.org/10.1080/02699206.2020.1734864